Scaling Econometrics: Text Processing, Distributed Computing, and Experimental Design

Yiğit Okar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Yanqin Fan, Chair

Jing Tao

Yuya Takahashi

Program Authorized to Offer Degree:

Economics

University of Washington

Abstract

Scaling Econometrics: Text Processing, Distributed Computing, and Experimental Design

Yiğit Okar

Chair of the Supervisory Committee:
Yanqin Fan
Economics

This dissertation develops new methodological approaches to address three fundamental challenges in modern econometrics: computational scalability in choice models, experimental design in digital markets, and the integration of unstructured text data. The first chapter addresses the computational challenges in estimating multinomial logistic regression models with large choice sets. We introduce an iterative distributed computing estimator that dramatically reduces computational burden while preserving statistical efficiency. This estimator, when initialized with a consistent preliminary estimate, achieves asymptotic efficiency under a weak dominance condition. We develop a parametric bootstrap procedure for statistical inference and establish its consistency. Through extensive simulation studies, we demonstrate that our method achieves substantial computational gains while maintaining estimation accuracy, making it particularly valuable for applications in industrial organization and marketing where researchers face increasingly large choice sets. The second chapter tackles the methodological challenges inherent in e-commerce pricing experiments. While cluster randomization is necessary to prevent bias from spillover effects between substitute products, it introduces additional variation that can compromise statistical power. We develop a comprehensive analytical framework for understanding and managing these variance components. Our methodology makes several contributions: first, we provide a detailed decomposition of variance components in cluster randomized experiments; second, we in-

troduce a novel binned estimator specifically designed for the high-kurtosis data common in e-commerce settings; and third, we evaluate various approaches to variance reduction including matched-pair designs, stratified randomization, and covariate adjustment. Through simulation of e-commerce data, we demonstrate that our proposed methods can improve power while maintaining robust inference. The binned estimator proves particularly effective, though we carefully describe the conditions under which it maintains unbiasedness. The third chapter presents a methodological breakthrough in the integration of textual data into econometric analysis. We develop a two-stage text regression methodology that leverages recent advances in transformer-based language models to capture rich semantic information and contextual nuances. The first stage employs state-of-the-art natural language processing techniques to represent textual data in a lower-dimensional space while preserving semantic relationships. The second stage develops an econometric framework for estimating the association between these text-derived features and economic outcomes. We demonstrate the methodology's effectiveness through an application to online economics forums, showing substantial improvements in both predictive accuracy and interpretability compared to traditional bag-of-words approaches. This methodology opens new avenues for research across various subfields of economics, from labor economics to finance, where textual data may provide crucial insights into economic behavior and outcomes. Collectively, these chapters advance the frontier of empirical methods in economics by developing scalable solutions for modern data challenges. The methodological innovations presented here enable researchers to handle larger datasets, conduct more precise experiments, and incorporate richer forms of information into their analyses. While each chapter addresses a distinct challenge, they are united by a common theme: expanding the scope of feasible empirical research through methodological innovation. The tools and frameworks developed in this dissertation contribute to the growing toolkit available to empirical researchers, particularly those working with large-scale, complex, or unstructured data.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I wish to express my profound gratitude to my advisor, Professor Yanqin Fan, whose extraordinary mentorship has shaped not only my dissertation but my entire approach to research. Her brilliant insights, rigorous scholarly standards, and unwavering support have been instrumental in my academic journey. Her dedication to excellence in research has set an example that will guide me throughout my career. I am truly fortunate to have had the opportunity to learn from such an exceptional scholar and mentor.

I am deeply grateful to my committee members, Professor Jing Tao and Professor Yuya Takahashi, for their insightful feedback and support.

I would like to express my sincere thanks to Jeff and Perri Roe for the fellowship they provided, which has been instrumental in supporting my research and academic pursuits, creating a compounding effect that continues to shape and elevate my career trajectory.

I would like to extend special thanks to Professor Xuetao Shi, whose calm demeanor, patience, and sharp intellectual insights have been truly inspiring. I am particularly indebted to Professor Tayfun Keskin at the Business School, who has been an exemplar of true academic scholarship. His practical approach to research and dedication to bringing academic work into the real world has deeply influenced my perspective.

Several faculty members have played crucial roles in my academic development. I am grateful to Professor Yu-chin Chen for her support as the graduate program director, and to Professor Marco Madunic for his guidance in academic communication. I would also like to express my heartfelt gratitude to Professor Burak Saltoglu from Boğaziçi University, whose unwavering belief in me and continued support since my undergraduate years have been invaluable. I am also deeply thankful to Professors Tolga Umut Kuzubas and Ayhan

Yuksel, also from Boğaziçi University, for their mentorship and encouragement throughout my academic journey.

The administrative staff have been instrumental in making my graduate life smoother. At the Foster Business School, I am thankful to Nuzulita Budhiari and Shannon Goodwin for their assistance. At the Economics Department, Simon Reeve-Parker, Heidi Hannah, and Michelle Foshee have provided invaluable support throughout different stages of my program.

My doctoral journey has been enriched by the friendships I've formed at UW. I am grateful to Sean Ewen, Jorge Rivero, Resem Makan, Raj Datta, John Kim, Seungryul Jeong, Amre Abken, Ryan Cummings, Dadmehr Didgar, Wendao Xue, Aochun Di, Abby Schamp, Reina Kawai, Kovid Puria, Yvonne Ng, Yoon Choi, and Hyeonseok Park for their support and friendship throughout these years. A special thank you to my dearest friend Bertan Kursun, whose friendship has made all the difference in both my life and career.

I wish to express my deepest gratitude to my grandparents and parents for their unconditional love and unwavering faith in me, which has been a constant source of motivation to improve and excel. To my twin brother, Mert Okar – you have been my constant inspiration and closest confidant throughout this journey. Your determination and success in your own path have always pushed me to be better, and your unwavering support has been a source of strength I could always count on. I dedicate this work to the memory of my uncle, Fatih Koseoglu, whom we lost during my doctoral studies. His belief in education and his pride in my academic journey continue to inspire me.

Finally, to my wife Nil – your patience, devotion, and understanding have been my anchor throughout this journey. Your love and care made it possible for me to complete this dissertation. This achievement is as much yours as it is mine. And to our beloved dog Bal, who has been my constant companion during countless hours of writing and research – your joyful presence and unconditional love have brightened even the most challenging days of this journey.

DEDICATION

to my dear wife, Nil & our princess, Bal

Chapter 1

ITERATIVE DISTRIBUTED MULTINOMIAL REGRESSION

1.1 Introduction

1.1.1 Motivation and Main Contributions

Discrete choice models, including logit and multinomial-logit (MNL) models, are widely used in applied social science research. With the growing availability of diverse data types and the integration of econometric models with textual and image data, researchers are encountering applications of MNL models with massive number of choices; see the applications discussed in Section 1.1.2. In these cases, even if the number of parameters for each choice is small, the total number of parameters which increases linearly with the number of choices will be large due to large choice sets. As a result, the maximum likelihood estimation (MLE) becomes computationally intractable due to the cost of solving the optimization problem for a large number of parameters.[1]

Researchers in diverse disciplines have explored various approaches and proposed numerous methods to numerically solve the MLE; see the numerical algorithms discussed in Section 1.1.2. However, there is a notable lack of theoretical results ensuring the consistency or asymptotic efficiency of the estimators derived from these methods. This gap in the literature between numerical computation and statistical inference motivates the present study.

This chapter is based on joint work with Yanqin Fan and Xuetao Shi.

[1]MLE of the MNL model refers to the conditional MLE given the covariate and total counts.

Our proposed estimator utilizes the multinomial-Poisson (MP) transformation, which reformulates the multinomial likelihood into a Poisson likelihood by incorporating individual fixed effects into the MNL model. *When all the covariates are categorical*, [Baker, 1994] establishes that the multinomial likelihood and the Poisson likelihood produce identical estimates of the parameters in the MNL model and advocates the computational advantage of maximizing the Poisson likelihood. However, the MP transformation has also been employed in MNL models with continuous or mixed discrete and continuous covariates. For instance, [Gentzkow et al., 2019b] note that they "... approximate the likelihood of [their] multinomial logit model with the likelihood of a Poisson model ...".

As the first contribution of this paper, we establish an equivalence result *for all types of covariates, discrete or mixed* justifying the application of the MP transformation in these cases. To accomplish this, we re-interpret the Poission likelihood from the MP transformation as a conditional quasi-log-likelihood function given the covariates. Maximizing this function provides a quasi-maximum likelihood estimator (QMLE) for the MNL model. The equivalence result is established as the equivalence between the resulting QMLE and MLE of the MNL.

While the QMLE is computationally more efficient than the MLE, it's still costly when applied to MNL models with large choice sets. To overcome this computational challenge, [Taddy, 2015a] exploits an important feature of the quasi-log-likelihood function: for any given fixed effects, the function is additively separable in the parameters across different choices of the MNL model. [Taddy, 2015a] proposes to estimate parameters for each choice separately at a specific value of the fixed effects and calls the resulting estimator *distributed computing* estimator. As noted in [Taddy, 2015a], however, his distributed computing estimator is inconsistent except in a few very special cases.

To regain consistency and asymptotic efficiency, we adopt the idea of iterative backfitting algorithms studied in [Pastorello et al., 2003], [Dominitz and Sherman, 2005], and [Fan et al., 2015] to compute QMLE of parameters in MNL and fixed effects iteratively.[2] Dur-

[2]Similar iterative algorithms have also been developed for dynamic discrete games of incomplete infor-

ing each iteration, we first solve for the parameters of interest through *distributed computing* given the approximate values of the individual fixed effects. Then we update estimates of the individual fixed effects based on their expressions from maximizing the quasi-log-likelihood function using the previous estimates of the model parameters. We call our estimator *iterative distributed computing* (IDC) estimator. This is the second and main contribution of the paper. The algorithm for IDC is fast to run because of distributed computing, even when the choice set is large. We consider three IDC estimators based on three different initial estimators: a consistent estimator based on pairwise binomial logistic regression, [Taddy, 2015a]'s distributed computing estimator, and a maximum likelihood estimator assuming that the distribution of total counts is Poisson. The latter two estimators are inconsistent in general. All three initial estimators are fast to compute because they allow for distributed computing.

As the third contribution, we establish theoretical result on the consistency and asymptotic efficiency of all three IDC estimators. When the initial estimator is consistent, we show that the IDC estimator with any finite number of iterations is always consistent and is asymptotically efficient under an information dominance condition when the number of iterations diverges with respect to the sample size n at $\log(n)$ rate. With inconsistent initial estimators, the IDC estimators are consistent and asymptotically efficient under a stronger contraction mapping condition when the number of iterations diverges with respect to n at polynomial rate.

When the number of choices is large, conducting inference via plug-in estimation of the variance matrix becomes infeasible. This is because the Fisher information matrix is of very large dimension, causing the computation of the inverse of its estimator both time-consuming and unreliable. The fourth contribution of the paper is that we propose a parametric bootstrap inference procedure and show its consistency. Because the IDC estimator is fast to compute, our inference procedure is computationally feasible.

---

mation where directly computing the MLE using the nested fixed-point algorithm proves computationally infeasible, see [**?**, **?**] for a nested pseudo-likelihood (NPL) algorithm. [**?**] further analyze the conditions necessary for the convergence of the NPL algorithm and derive its convergence rate.

Lastly, we conduct extensive simulations to study the finite sample performance of our estimator and inference procedure. We are particularly interested in the computational time of the IDC estimator and its accuracy in comparison with the maximum likelihood estimator. The simulation results show that the IDC estimator is very fast to compute with its running time being approximately linear in the number of choices. Compared to the maximum likelihood estimator, our estimator has a very similar mean squared error in all the different model settings but is much faster to compute when the number of choices is large. We also study the finite sample behavior of the proposed bootstrap inference procedure. The result suggests that the procedure achieves the correct size and is consistent.

1.1.2 Related Literature

Applications The proposed IDC estimator can be applied to study various economics and computer science topics such as text analysis, dimensionality reduction, spatial choice models, image classification ([Russakovsky et al., 2015]), and video recommendation ([Davidson et al., 2010]).

Text analysis: The integration of text data into econometric models is increasingly prominent in economics. For example, [Baker and Wurgler, 2006a] analyze investor sentiment's effect on stock returns, while [Chen et al., 2021] explore how hedge funds capitalize on sentiment changes. Modeling text data often involves treating word counts as a multinomial distribution, as [Taddy, 2015a] demonstrates using Yelp reviews to predict outcomes based on user and business attributes. [Gentzkow et al., 2019b] use distributed computing estimator for the multinomial regression to measure polarization in Congressional speeches. [Kelly et al., 2019] extend this approach, using Hurdle Distributed Multinomial Regression to backcast, nowcast, and forecast macroeconomic variables from newspaper text.

Dimensionality Reduction: Our estimator aids in dimensionality reduction for inverse multinomial regression, as [Taddy, 2013] discusses. Instead of inferring sentiment from text, [Taddy, 2013]'s approach estimates word distribution given sentiment. He introduces a score based on word frequencies and regression parameters, useful in forward-regression models.

Spatial Choice Models: High-dimensional choices also appear in spatial models. [Buchholz, 2021]

models taxi drivers' location choices with a dynamic spatial search, reducing dimensionality via discretization. Similarly, [Pellegrini and Fotheringham, 2002] apply hierarchical discrete choice models to immigration, while [Bettman, 1979] addresses brand choices in limited-option settings, proposing hierarchical selection for high-dimensional cases.

Numerical Algorithms To address the computational difficulty of solving the MLE of the MNL model, researchers have proposed several numerical methods to find approximate solutions. [Böhning and Lindsay, 1988] and [Böhning, 1992] propose to replace the Hessian matrix in the Newton-Raphson iteration with its easy-to-compute global lower bound and show that the approximate solution converges with the number of iterations. Because the convergence rate depends crucially on the difference between the Hessian matrix and its lower bound, the algorithm can be slow to run for certain model parameters. Additionally, based on our simulation exercise, if the choice probabilities vary significantly across different choices with some being close to zero, the algorithm becomes unstable. In comparison, our IDC algorithm is stable in all the simulation settings. [Boyd et al., 2011] introduce an alternating direction method of multipliers, which reformulates the original optimization problem by introducing redundant linear constraints. [Gopal and Yang, 2013] propose a log concavity method, which replaces the log partition function of the multinomial logit with a parallelizable upper-bound. [Recht et al., 2011], [Raman et al., 2016], and [Fagan and Iyengar, 2018] study a stochastic gradient descent method, which uses random training samples to calculate the gradient at each iteration. Although these methods can be computationally efficient, to the best of the authors' knowledge, no consistency or asymptotic efficiency result has been shown in these work.

Penalization methods have also been introduced to the MNL regression and some of the numerical methods discussed above are adopted in solving the penalized MNL regression, see e.g., [Friedman et al., 2010], [Simon et al., 2013], and [Nibbering and Hastie, 2022], The proposed IDC procedure in this paper can be combined with the aforementioned algorithms to further improve the performance of penalized MNL regressions.

Organization of the rest of this paper The remainder of this paper is organized as follows. In Section 1.2 we present a comprehensive overview of the multinomial logistic regression model and the MP transformation. Section 1.3 introduces our iterative distributed computing estimator along with some initial values. In Section 1.4 we provide the asymptotic theory of the iterative distributed computing estimator. Section 1.5 contains the simulation results. Finally with Section 1.6 we conclude. Appendix A.1 collects the notations and equations used in the paper. All the technical proofs are provided in Appendix A.2. The codes for implementing the estimation and inference procedures are available at here.

Notations Throughout the paper, we use index $i \in \{1, \ldots, n\}$ for individual, $j \in \{1, \ldots, p\}$ for covariate, and $k \in \{1, \ldots, d\}$ for unique choice. Boldfaced symbols such as $\boldsymbol{C}$ and $\boldsymbol{V}$ are used to denote vectors; while elements of the vectors are denoted by plain symbol such as C_k and V_j. Denote $\sim$ as 'equality up to a constant", such that $f(\theta) \sim g(\theta)$ is equivalent to $f(\theta) = g(\theta) + h$, where h is a constant relative to θ.

1.2 Multinomial Logistic Regression

Let $\boldsymbol{C}_i \in \mathbb{R}^d$ denote the random vector of counts on d different choices for individual $i = 1, ..., n$, summing up to $M_i = \sum_{k=1}^{d} C_{ik}$. We use the random vector $\boldsymbol{V}_i \in \mathbb{R}^p$ to denote the covariate vector that includes a constant.

Consider a *correctly specified* multinomial-logit (MNL) model. The conditional probability mass function is given by the following:

$$\Pr(\boldsymbol{C}_i \mid \boldsymbol{V}_i, M_i) = \mathrm{MNL}(\boldsymbol{C}_i; \boldsymbol{\eta}_i^*, M_i) \equiv \frac{M_i!}{C_{i1}! \cdots C_{id}!} \left(\frac{e^{\eta_{i1}^*}}{\Lambda_i^*}\right)^{C_{i1}} \cdots \left(\frac{e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{C_{id}}, \tag{1.2.1}$$

where for $k = 1, \ldots d$, we let $\eta_{ik}^* \equiv \boldsymbol{V}_i' \boldsymbol{\theta}_k^*$ with unknown parameters $\boldsymbol{\theta}_k^* \equiv \left(\theta_{k1}^*, \ldots, \theta_{kp}^*\right)'$ and $\Lambda_i^* \equiv \sum_{k=1}^{d} e^{\eta_{ik}^*}$. For the identification, we set $\boldsymbol{\theta}_d^* = \mathbf{0}$. Let $\boldsymbol{\theta}^* \equiv (\boldsymbol{\theta}_1^{*\prime}, \ldots, \boldsymbol{\theta}_d^{*\prime})' \in \Theta$ denote the parameter vector of interest. Throughout the paper, we use the supercript $*$ to indicate the true value of the unknown parameter. Denote the parameter space of $\boldsymbol{\theta}_k^*$ for $k = 1, \ldots, d$

as Θ_k. We have that $\Theta_d = \{\mathbf{0}\}$ and $\Theta \equiv \prod_{k=1}^{d} \Theta_k$.

In this paper, we focus on the case where d is large (but fixed) such that directly solving for the maximum likelihood estimator is computationally costly. Applications include text corpora, where $\boldsymbol{C}_i$ represents the counts of d different words/phrases in a text of M_i words; browser logs, where $\boldsymbol{C}_i$ indicates the number of times a website among d total websites is visited by an individual; and location choices, where among M_i number of locations traveled by the driver, $\boldsymbol{C}_i$ contains the number of times each location, among d different ones, is visited.

1.2.1 Maximum Likelihood Estimation (MLE)

Given a random sample of size n, let $\boldsymbol{\eta}_i \equiv (\eta_{i1}, \ldots, \eta_{id})' \equiv (\boldsymbol{V}_i'\boldsymbol{\theta}_1, \ldots, \boldsymbol{V}_i'\boldsymbol{\theta}_d)$ and $\Lambda_i = \sum_{k=1}^{d} e^{\eta_{ik}}$ for $i = 1, ..., n$. Ignoring terms that are independent of the parameter $\boldsymbol{\theta}$, the conditional log-likelihood function given the covariate $\boldsymbol{V}$ and total count M takes the following form:

$$
\begin{aligned}
l_{C|V,M}(\boldsymbol{\theta}) &\equiv \sum_{i=1}^{n} \log \Pr(\boldsymbol{C}_i \mid \boldsymbol{V}_i, M_i) \\
&\sim \sum_{i=1}^{n} \log \left[\left(\frac{e^{\eta_{i1}}}{\Lambda_i} \right)^{C_{i1}} \cdots \left(\frac{e^{\eta_{id}}}{\Lambda_i} \right)^{C_{id}} \right] \\
&= \sum_{i=1}^{n} \left\{ C_{i1} \left[\log\left(e^{\eta_{i1}}\right) - \log\left(\sum_{k=1}^{d} e^{\eta_{ik}} \right) \right] + \cdots + C_{id} \left[\log\left(e^{\eta_{id}}\right) - \log\left(\sum_{k=1}^{d} e^{\eta_{ik}} \right) \right] \right\} \\
&= \sum_{i=1}^{n} \left[C_{i1}\eta_{i1} + \cdots + C_{id}\eta_{id} - (C_{i1} + \cdots + C_{id}) \log\left(\sum_{k=1}^{d} e^{\eta_{ik}} \right) \right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{C}_i'\boldsymbol{\eta}_i - M_i \log\left(\sum_{k=1}^{d} e^{\eta_{ik}} \right) \right]. \qquad (1.2.2)
\end{aligned}
$$

Let $L_{C|V,M}(\boldsymbol{\theta})$ denote the probability limit of $\frac{1}{n} l_{C|V,M}(\boldsymbol{\theta})$. Denote $B(\boldsymbol{\theta}, \varepsilon)$ as an open ball in Θ centered at $\boldsymbol{\theta}$ with radius ε. We make the following assumption throughout the paper.

Assumption 1.2.1. *(i) The true value* $\boldsymbol{\theta}^* \in \Theta$ *satisfies that* $\sup_{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}^*, \varepsilon)} L_{C|V,M}(\boldsymbol{\theta}) <$

$L_{C|V,M}(\boldsymbol{\theta}^*)$ *for any* $\varepsilon > 0$. *(ii)* $\boldsymbol{\theta}^*$ *is in the interior of* Θ.

Assumption 1.2.1 (i) implies that $\boldsymbol{\theta}^*$ is identified as $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\Theta} L_{C|V,M}(\boldsymbol{\theta})$. Define the following objective function:

$$Q_n^*(\boldsymbol{\theta}) \equiv -l_{C|V,M}(\boldsymbol{\theta}). \tag{1.2.3}$$

Based on (1.2.3), the conditional maximum likelihood estimator of $\boldsymbol{\theta}^*$ is:[3]

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*(\boldsymbol{\theta}). \tag{1.2.4}$$

Solving the above optimization problem analytically is impossible. In addition, due to the potentially large dimension d, numerical algorithms such as the Newton-Raphson method are difficult to implement either because they usually involve computing the inverse of the Hessian matrix, which is of dimension $pd \times pd$, during each iteration. In this paper, we propose an estimator that is both computationally attractive and asymptotically efficient.

1.2.2 Multinomial-Poisson Transformation

In this section, we present the multinomial-Poisson (MP) transformation, based on which we develop our estimator. We reinterpret the Poission likelihood as a quasi-likelihood conditional on the covariates.

Let $\mathbf{1}_d \equiv (1,\ldots,1)' \in \mathbb{R}^d$. With slight abuse of notation, define

$$\begin{aligned} l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu}) &\equiv l_{C|V,M}(\boldsymbol{\theta}) + \sum_{i=1}^{n} \left[\mu_i \boldsymbol{C}_i' \mathbf{1}_d - M_i \log\left(e^{\mu_i}\right)\right] \\ &= \sum_{i=1}^{n} \left[\boldsymbol{C}_i'(\boldsymbol{\eta}_i + \mu_i \mathbf{1}_d) - M_i \log\left(\sum_{k=1}^{d} e^{\eta_{ik}+\mu_i}\right)\right], \end{aligned}$$

[3]The definition of $\widetilde{\boldsymbol{\theta}}$ implicitly assumes that the solution to the minimization problem is unique. This can be shown to hold with probability approaching one by the identification of the model. See [McFadden, 1973]. The same result holds for all the estimators defined in the paper. We ignore such mathematical subtlety for the remainder of the paper to simplify the discussion.

where $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$. The following lemma shows that the two functions $l_{C|V,M}(\boldsymbol{\theta})$ and $l_{C|V,M}(\boldsymbol{\theta}, \boldsymbol{\mu})$ are the same for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. In other words, argument $\boldsymbol{\mu}$ in $l_{C|V,M}(\boldsymbol{\theta}, \boldsymbol{\mu})$ does not affect the value of the function. The proof of the lemma is straightforward by realizing that $\boldsymbol{C}_i' \mathbf{1}_d = M_i$ by definition.

Lemma 1.2.1. *$l_{C|V,M}(\boldsymbol{\theta}) = l_{C|V,M}(\boldsymbol{\theta}, \boldsymbol{\mu})$ for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\mu} \in \mathbb{R}^n$.*

Define the following two functions:

$$f(\boldsymbol{\theta}, \boldsymbol{\mu}) \equiv \sum_{i=1}^{n} \left[M_i \log \left(\sum_{k=1}^{d} e^{\eta_{ik} + \mu_i} \right) - \sum_{k=1}^{d} e^{\eta_{ik} + \mu_i} \right] \text{ and} \tag{1.2.5}$$

$$\begin{aligned} ql_{C|V}(\boldsymbol{\theta}, \boldsymbol{\mu}) &\equiv l_{C|V,M}(\boldsymbol{\theta}, \boldsymbol{\mu}) + f(\boldsymbol{\theta}, \boldsymbol{\mu}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{d} \left(C_{ik} (\eta_{ik} + \mu_i) - e^{(\eta_{ik} + \mu_i)} \right). \end{aligned} \tag{1.2.6}$$

It is not difficult to see that $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ takes the form of a log-likelihood function of n conditional Poisson distributions with means $\sum_{k=1}^{d} e^{(\eta_{ik} + \mu_i)}$, $i = 1, \ldots, n$ (after ignoring terms that are independent of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$). This in turn renders $ql_{C|V}$ a conditional quasi-log-likelihood function of which C_{ik} given V_i is drawn independently from a Poisson distribution with mean $e^{(\eta_{ik} + \mu_i)}$, $k = 1, ..., d$. This property underlies the naming of the MP transformation.

Based on the conditional quasi-log-likelihood function $ql_{C|V}(\boldsymbol{\theta}, \boldsymbol{\mu})$, we can compute a conditional quasi MLE (QMLE) of $\boldsymbol{\theta}^*$:

$$\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\mu}} \right) = \arg \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\mu} \in \mathbb{R}^n} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu}), \tag{1.2.7}$$

$$\text{where } Q_n(\boldsymbol{\theta}, \boldsymbol{\mu}) \equiv -ql_{C|V}(\boldsymbol{\theta}, \boldsymbol{\mu}).$$

[Baker, 1994] shows that $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$ if the covariate vector $\boldsymbol{V}$ contains only categorical random variables. The following lemma demonstrates that $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$ holds irrespective of the type of the covariate vector, thereby generalizing the result of [Baker, 1994].

Lemma 1.2.2. *It holds that $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$.*

It is worth noting that Lemma 1.2.2 does not rely on the assumption that $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ is the correct log-likelihood function of M_i or equivalently $\Pr(M_i \mid \boldsymbol{V}_i) = \text{Po}\left(\sum_{k=1}^{d} e^{(\eta_{ik}+\mu_i)}\right)$ or $\Pr(C_{ik} \mid \boldsymbol{V}_i) = \text{Po}\left(e^{(\eta_{ik}+\mu_i)}\right)$, where $\text{Po}(\cdot)$ denotes the Poisson distribution. No assumption on the conditional distribution of M_i given $\boldsymbol{V}_i$ is needed for any of the results in the paper to hold. As we show in the following sections, introducing $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ is merely a trick to achieve distributed computing.

1.3 Iterative Distributed Computing Estimator

Lemma 1.2.2 shows that instead of minimizing $Q_n^*(\boldsymbol{\theta})$, we can minimize $Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ to obtain the QMLE of $\boldsymbol{\theta}^*$. However, the computation of $\widehat{\boldsymbol{\theta}}$ is also time-consuming, since solving (1.2.7) is infeasible in practice if d is large. However the additive form of $Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ allows solving (1.2.7) distributively.

1.3.1 Distributed Computing Estimator in [Taddy, 2015a]

As noted in [Taddy, 2015a], although minimizing $Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ jointly is computationally infeasible, given any value of $\boldsymbol{\mu}$, numerically solving $\arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ is much easier because the optimization can be done separately for each $\boldsymbol{\theta}_k$ and be computed across machines. To see this, we rewrite $Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ as:

$$\begin{aligned} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \sum_{i=1}^{n}\sum_{k=1}^{d}\left(e^{(\eta_{ik}+\mu_i)} - C_{ik}(\eta_{ik}+\mu_i)\right). \\ &= \sum_{k=1}^{d}\sum_{i=1}^{n}\left(e^{\left(\boldsymbol{V}_i'\boldsymbol{\theta}_k+\mu_i\right)} - C_{ik}(V_i'\boldsymbol{\theta}_k+\mu_i)\right) \\ &\equiv Q_{1n}(\boldsymbol{\theta}_1, \boldsymbol{\mu}) + \cdots + Q_{dn}(\boldsymbol{\theta}_d, \boldsymbol{\mu}), \end{aligned} \tag{1.3.1}$$

where for $k = 1, \ldots, d$, $Q_{kn}(\boldsymbol{\theta}_k, \boldsymbol{\mu}) \equiv \sum_{i=1}^{n} \left(e^{(V_i' \boldsymbol{\theta}_k + \mu_i)} - C_{ik} (V_i' \boldsymbol{\theta}_k + \mu_i) \right)$. In consequence, it holds that for any $\boldsymbol{\mu}$,

$$\arg\min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu}) = \left[\arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} Q_{1n}(\boldsymbol{\theta}_1, \boldsymbol{\mu}), \ldots, \arg\min_{\boldsymbol{\theta}_d \in \Theta_d} Q_{dn}(\boldsymbol{\theta}_d, \boldsymbol{\mu}) \right]'. \tag{1.3.2}$$

Based on (1.3.2), solving $\arg\min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ for any given $\boldsymbol{\mu}$ is equivalent to solving d optimizations: $\arg\min_{\boldsymbol{\theta}_k \in \Theta_k} Q_{kn}(\boldsymbol{\theta}_k, \boldsymbol{\mu})$ for each $k = 1, \ldots, d$, where each optimization is a Poisson regression.[4] Since $\boldsymbol{\theta}_k$ has only p dimensions, $\arg\min_{\boldsymbol{\theta}_k \in \Theta_k} Q_{kn}(\boldsymbol{\theta}_k, \boldsymbol{\mu})$ is easy to compute. In addition, the optimizations for $k = 1, \ldots, d$ can be computed across machines allowing for distributed computing.

By Equation (1.2.7), it is not difficult to see that $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}})$. Because $\widehat{\boldsymbol{\theta}}$ is equivalent to the MLE $\widetilde{\boldsymbol{\theta}}$ by Lemma 1.2.2, it has the desired properties such as being both consistent and asymptotically efficient. As a result, we would hope to obtain $\widehat{\boldsymbol{\mu}}$ first and then compute $\widehat{\boldsymbol{\theta}}$ by distributed computing. However, the value of $\widehat{\boldsymbol{\mu}}$ depends on $\widehat{\boldsymbol{\theta}}$, which itself is difficult to calculate. On the other hand, given any value of $\boldsymbol{\theta}$, solving $\arg\min_{\boldsymbol{\mu}} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ is also fast, and the solution even has a closed form. Denote the solution to $\arg\min_{\boldsymbol{\mu}} Q_n(\boldsymbol{\theta}, \boldsymbol{\mu})$ as $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$. Simple calculation would show that

$$\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta}) = \left(\log\left(\frac{M_1}{\sum_{k=1}^{d} e^{\eta_{1k}}} \right), \ldots, \log\left(\frac{M_n}{\sum_{k=1}^{d} e^{\eta_{nk}}} \right) \right)'. \tag{1.3.3}$$

Let $\widehat{\boldsymbol{\mu}}_T = (\log(M_1), \ldots, \log(M_n))'$. Instead of solving for $\widehat{\boldsymbol{\mu}}$ using (1.3.3), [Taddy, 2015a] proposes an estimator $\widehat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_T)$ and calls it the distributed computing estimator. Such an estimator is fast to compute but fails to be consistent except in the special cases discussed in [Taddy, 2015a].

[4] Since $\Theta_d = \{\mathbf{0}\}$, solving for $\arg\min_{\boldsymbol{\theta}_d \in \Theta_d} Q_{dn}(\boldsymbol{\theta}_d, \boldsymbol{\mu})$ is trivial.

1.3.2 Iterative Distributed Computing Estimator

We propose an iterative distributed computing (IDC) estimator, such that during each iteration we solve (1.3.2) with $\boldsymbol{\mu}$ updated from the previous step estimate of $\boldsymbol{\theta}$ via (1.3.3). Our IDC estimator is defined by the following steps.

Step 0. Compute an initial estimator of $\boldsymbol{\theta}^*$, denoted as $\widehat{\boldsymbol{\theta}}^{(0)}$.

Step 1, . . . , **S.** For step s, where $s = 1, \ldots, S$, we first update $\boldsymbol{\mu}$ using estimator $\widehat{\boldsymbol{\theta}}^{(s-1)}$ from the previous step via $\overline{\boldsymbol{\mu}}_n(\cdot)$. Then we update $\boldsymbol{\theta}$ given the value of $\boldsymbol{\mu}$:

$$\begin{aligned}
&\widehat{\boldsymbol{\theta}}^{(s)} \equiv \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right) \\
&= \left[\arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right), \ldots, \arg\min_{\boldsymbol{\theta}_d \in \Theta_d} Q_{dn}\left(\boldsymbol{\theta}_d, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)\right]'. \qquad (1.3.4)
\end{aligned}$$

The iterative estimator with S iterations is defined as $\widehat{\boldsymbol{\theta}}^{I} \equiv \widehat{\boldsymbol{\theta}}^{(S)}$. For any $\boldsymbol{\theta}$, the value of $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$ can be directly computed from (1.3.3). In each step, we compute $\arg\min_{\boldsymbol{\theta}_k} Q_{kn}\left(\boldsymbol{\theta}_k, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)$ for $k = 1, \ldots, d$ on d parallel computers. This amounts to running d Poisson regressions with p parameters, and the computational burden for each step is low. The algorithm is described in Algorithm 1.

Unlike many existing algorithms such as gradient descent or stochastic gradient descent, the IDC estimator does not involve any tuning parameter. This is advantageous because the performance of the classical (stochastic) gradient descent is generally sensitive to the learning rate.

1.3.3 Initial Estimators

Similar to all iterative optimization procedures, the initial estimator plays a critical role. In finite samples, a good initial guess of $\boldsymbol{\theta}^*$ can improve the performance of the IDC estimator. Asymptotically, a consistent initial estimator can lead to consistent and asymptotically efficient iterative estimators under weaker assumptions than an inconsistent initial estimator. In this section, we propose three initial estimators: a consistent initial estimator of $\boldsymbol{\theta}^*$ based on

Algorithm 1: the iterative distributed computing procedure

Input: S

Output: $\widehat{\boldsymbol{\theta}}^{(S)}$

Compute an initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$

/* Start of Step 1 */

Compute $\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)$

Solve for

$$\widehat{\boldsymbol{\theta}}^{(1)} = \left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right), \ldots, \arg\min_{\boldsymbol{\theta}_d\in\Theta_d} Q_{dn}\left(\boldsymbol{\theta}_d, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right)\right]'$$

/* End of Step 1. The output is $\widehat{\boldsymbol{\theta}}^{(1)}$ */

/* Start of Step 2 */

Compute $\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(1)}\right)$

Solve for

$$\widehat{\boldsymbol{\theta}}^{(2)} = \left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(1)}\right)\right), \ldots, \arg\min_{\boldsymbol{\theta}_d\in\Theta_d} Q_{dn}\left(\boldsymbol{\theta}_d, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(1)}\right)\right)\right]'$$

/* End of Step 2. The output is $\widehat{\boldsymbol{\theta}}^{(2)}$ */

..... Continue until Step S. The output of Step S is $\widehat{\boldsymbol{\theta}}^{(S)}$

binomial MLE, [Taddy, 2015a]'s estimator, and the MLE based on the Poisson assumption of M_i. The latter two are inconsistent without any assumption on the distribution of M_i.

A Consistent Initial Estimator Let $N_{ik} \equiv C_{ik} + C_{id}$. The following lemma results from the MNL model defined in (1.2.1).

Lemma 1.3.1. *For any* $k = 1, \ldots, d-1$,

$$\Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, N_{ik}\right) = \frac{N_{ik}!}{C_{ik}!C_{id}!}\left(\frac{e^{\eta_{ik}^*}}{e^{\eta_{ik}^*}+1}\right)^{C_{ik}}\left(\frac{1}{e^{\eta_{ik}^*}+1}\right)^{C_{id}}. \tag{1.3.5}$$

Lemma 1.3.1 shows that we can consistently estimate $\boldsymbol{\theta}_k^*$ based on a binomial logistic

regression with the log-likelihood function given by:

$$\begin{aligned} l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right) &\equiv \sum_{i=1}^{n} \log \Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, N_{ik}\right) \\ &\sim \sum_{i=1}^{n} \left[C_{ik}\eta_{ik} - \left(C_{ik}+C_{id}\right)\log\left(e^{\eta_{ik}}+1\right)\right] \\ &= \sum_{i=1}^{n} \left[C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k - \left(C_{ik}+C_{id}\right)\log\left(e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}+1\right)\right]. \end{aligned}$$

Let $\check{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k \in \Theta_k} -l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right)$ for $k = 1, \ldots, d-1$ and $\check{\boldsymbol{\theta}} = \left(\check{\boldsymbol{\theta}}_1', \ldots, \check{\boldsymbol{\theta}}_{d-1}', \check{\boldsymbol{\theta}}_d'\right)'$ with $\check{\boldsymbol{\theta}}_d' = \mathbf{0}$. The consistency of $\check{\boldsymbol{\theta}}$ follows from the standard argument in the maximum likelihood estimation.

Compared to $\widetilde{\boldsymbol{\theta}}$, the conditional probability used in constructing the above binomial logistic log-likelihood function does not use all the available information. Therefore, $\check{\boldsymbol{\theta}}$ is less efficient than $\widetilde{\boldsymbol{\theta}}$. However, each component of $\check{\boldsymbol{\theta}}$, $\check{\boldsymbol{\theta}}_k$, can be calculated independently, allowing for parallel computing. The substantially short running time of $\check{\boldsymbol{\theta}}$ makes it a great candidate for the initial value $\widehat{\boldsymbol{\theta}}^{(0)}$.

Inconsistent Initial Estimators Even though [Taddy, 2015a]'s estimator fails to be consistent in general, it could serve as a candidate for the initial value in our Algorithm 1. Another option is to replace $\widehat{\boldsymbol{\mu}}_T$ with a zero vector to obtain another estimator denoted as $\widehat{\boldsymbol{\theta}}_P = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \mathbf{0}\right)$. It can also be computed across machines for each $k = 1, ..., d$. Like [Taddy, 2015a]'s estimator, $\widehat{\boldsymbol{\theta}}_P$ could also serve as a candidate for the initial value in our algorithm. Moreover, under an extra condition that M_i follows a Poisson distribution, $\widehat{\boldsymbol{\theta}}_P$ is the maximum likelihood estimator of $\boldsymbol{\theta}^*$.

Lemma 1.3.2. *If* $\Pr\left(M_i \mid V_i\right) = \mathrm{Po}\left(\sum_{k=1}^{d} e^{\eta_{ik}^*}\right)$, *then* $\widehat{\boldsymbol{\theta}}_P$ *is the maximum likelihood estimator of* $\boldsymbol{\theta}^*$ *based on the conditional probability* $\Pr\left(\boldsymbol{C}_i \mid V_i\right)$.

Unlike $\check{\boldsymbol{\theta}}$, neither $\widehat{\boldsymbol{\theta}}_T$ nor $\widehat{\boldsymbol{\theta}}_P$ is a consistent estimator of $\boldsymbol{\theta}^*$ without any additional assumption.

1.3.4 Constrained Iterative Distributed Computing

In some applications, researchers may have prior knowledge on some linear equality constraints among parameters. Taking the constraints into consideration during the estimation would further improve asymptotic efficiency. In this section, we discuss how to modify our IDC estimator introduced in Section 1.3.2 to incorporate equality constraints. The initial estimators introduced in Section 1.3.3, although they do not account for the equality constraints, can be utilized to obtain the initial $\widehat{\boldsymbol{\theta}}^{(0)}$.

We consider two different types of constraints: constraints on parameters for the same choice and for different choices. The procedures for different types of constraints differ in the optimization problems during each iteration. For each type, we use an example to illustrate our procedure.

For the first type, the constraint is on components of individual $\boldsymbol{\theta}_k^*$. Take the constraint $\theta_{k1}^* = \theta_{k2}^*$ for $k = 1, \dots d$ as an example. When computing $\widehat{\boldsymbol{\theta}}^{(s)}$ in Step s, we solve the constrained optimization problem:

$$\arg \min_{\boldsymbol{\theta}_k \in \Theta_k, \theta_{k1} = \theta_{k2}} Q_{kn}\left(\boldsymbol{\theta}_k, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)$$

for each k. Because the constraint is on each $\boldsymbol{\theta}_k^*$, the original distributed computing scheme remains.

The second type of constraints involves components of $\boldsymbol{\theta}_k^*$ across different choices. For example, researchers may impose restrictions like $\theta_{11}^* = \cdots = \theta_{q1}^*$, where $q < d$. For $k = 1, \dots, q$, let $\boldsymbol{\theta}_{k,-1}$ be the subvector of $\boldsymbol{\theta}_k$ that excludes its first element. From Steps 1 to S, we first update $\boldsymbol{\mu}$ using estimator $\widehat{\boldsymbol{\theta}}^{(s-1)}$ from the previous step. Then, we compute $\widehat{\boldsymbol{\theta}}^{(s)}$ from

the following optimization problems:

$$\begin{aligned}
&\left(\widehat{\boldsymbol{\theta}}_{1,-1}^{(s)}, \ldots, \widehat{\boldsymbol{\theta}}_{q,-1}^{(s)}, \widehat{\boldsymbol{\theta}}_{q+1}^{(s)}, \ldots, \widehat{\boldsymbol{\theta}}_{d}^{(s)}\right) \\
=&\Bigg[\arg\min_{\boldsymbol{\theta}_{1,-1}} Q_{1n}\left(\widehat{\theta}_{11}^{(s-1)}, \boldsymbol{\theta}_{1,-1}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right), \ldots, \arg\min_{\boldsymbol{\theta}_{q,-1}} Q_{qn}\left(\widehat{\theta}_{q1}^{(s-1)}, \boldsymbol{\theta}_{q,-1}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right) \\
&\arg\min_{\boldsymbol{\theta}_{q+1}} Q_{(q+1)n}\left(\boldsymbol{\theta}_{q+1}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right), \ldots, \arg\min_{\boldsymbol{\theta}_{d}} Q_{dn}\left(\boldsymbol{\theta}_{d}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)\Bigg], \qquad (1.3.6)
\end{aligned}$$

$$\widehat{\boldsymbol{\theta}}_{11}^{(s)} = \arg\min_{\theta_{11}}\left[Q_{1n}\left(\theta_{11}, \widehat{\boldsymbol{\theta}}_{1,-1}^{(s)}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right) + \cdots + Q_{qn}\left(\theta_{11}, \widehat{\boldsymbol{\theta}}_{q,-1}^{(s)}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)\right], \qquad (1.3.7)$$

and $\left(\widehat{\boldsymbol{\theta}}_{21}^{(s)}, \ldots, \widehat{\boldsymbol{\theta}}_{q1}^{(s)}\right) = \left(\widehat{\boldsymbol{\theta}}_{11}^{(s)}, \ldots, \widehat{\boldsymbol{\theta}}_{11}^{(s)}\right)$.

Optimization problems (1.3.6) can be solved using parallel computers. And (1.3.7) is an optimization problem with only one argument.[5] In consequence, each step incurs a low computational burden. The IDC estimator with such a constraint is also fast to compute.

The aforementioned two procedures can be generalized to accommodate any linear equality constraint. In particular, the two procedures can be combined in a straightforward way when constraints contain both types.

1.4 Asymptotic Theory

In this section, we establish the consistency and asymptotic normality of our IDC estimator introduced in Section 1.3.2.[6] Technical proofs are collected in Appendix A.2. We first impose the following two assumptions.

---

[5]Let a general linear equality constraint of the second type be written as $R\boldsymbol{\theta}^* = r$, where R and r are known with dimensions $l_R \times pd$ and $l_R \times 1$ respectively. The matrix R is assumed to have full row rank so that there is no redundant constraint. We can always rearrange and decompose R as $[R_c, \mathbf{0}]$, where R_c has dimension $l_R \times q$ and has no zero column. The number of arguments in the optimization problem (1.3.7) for $R\boldsymbol{\theta}^* = r$ is $q - l_R \geq 0$. The case where $q = l_R$ corresponds all q number of elements in $\boldsymbol{\theta}^*$ having prespecified values. The optimization problem (1.3.7) is no longer needed in this case.

[6]Asymptotic properties of the constrained estimators discussed in Section 1.3.4 can be established in the same way. Moreover, the procedures can be used to construct a likelihood ratio test for testing the null hypothesis on linear equality constraint among parameters.

Assumption 1.4.1. $\{(\boldsymbol{C}_i, \boldsymbol{V}_i, M_i)\}_{i=1}^n$ *are random samples of* $(\boldsymbol{C}, \boldsymbol{V}, M)$.

Assumption 1.4.2. *(i)* Θ *is compact and convex. (ii)* $\mathbb{E}\left[e^{\boldsymbol{V}'\boldsymbol{\theta}}\right] < \infty$ *for all* $\boldsymbol{\theta} \in \Theta$. *(iii)* $\mathbb{E}\left[M^2\right] < \infty$.

Assumption 1.4.2 is standard. Assumption 1.4.2 (ii) requires that the moment generating function of $\boldsymbol{V}$ exists within Θ. Almost all commonly seen distributions satisfy Assumption 1.4.2 (ii) and (iii).

1.4.1 Consistency of the IDC Estimator

In order to analyze the asymptotic properties of the IDC estimator, we need to study the way in which the iterative estimator updates itself in each step. In Step s of the iteration, we first compute $\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)$ using $\widehat{\boldsymbol{\theta}}^{(s-1)}$ from the previous step and then calculate $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)$. To explicitly distinguish the argument in $\overline{\boldsymbol{\mu}}_n(\cdot)$ from the argument in $Q_n(\cdot, \boldsymbol{\mu})$ for any given $\boldsymbol{\mu}$, we introduce $\boldsymbol{\vartheta}$ and use it as the argument in $\overline{\boldsymbol{\mu}}_n(\cdot)$. Define $Q_n^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \equiv Q_n(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n(\boldsymbol{\vartheta}))$. We have that

$$Q_n^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \equiv \sum_{i=1}^n \sum_{k=1}^d \left(\frac{M_i e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}}{\sum_{k=1}^d e^{\boldsymbol{V}_i'\boldsymbol{\vartheta}_k}} - C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k - C_{ik}\log\left(\frac{M_i}{\sum_{k=1}^d e^{\boldsymbol{V}_i'\boldsymbol{\vartheta}_k}}\right)\right).$$

Further define function $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ as the probability limit of $\frac{1}{n}Q_n^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$:

$$Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \equiv \sum_{k=1}^d \mathbb{E}\left[\left(\frac{M e^{\boldsymbol{V}'\boldsymbol{\theta}_k}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k - C_k\log\left(\frac{M}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\right)\right)\right]. \tag{1.4.1}$$

In the following lemma, we provide some properties of $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, which are crucial for obtaining the consistency and asymptotic normality of our IDC estimator.

Lemma 1.4.1. *Under Assumptions 1.4.1 and 1.4.2, the following results hold.*

(i) $\sup_{\boldsymbol{\theta},\boldsymbol{\vartheta}\in\Theta}\left|Q_n^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta}) - Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})\right| \xrightarrow{p} 0$.

(ii) For any given $\boldsymbol{\vartheta}$, $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ *has a unique minimizer denoted as* $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$.

(iii) $\overline{\boldsymbol{\theta}}(\cdot)$ *is continuous on* Θ.

(iv) $\overline{\boldsymbol{\theta}}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$, *i.e., the true value* $\boldsymbol{\theta}^*$ *is a fixed point of the mapping* $\overline{\boldsymbol{\theta}}: \Theta \to \Theta$.

(v) $\boldsymbol{\theta}^*$ *is the unique fixed point of* $\overline{\boldsymbol{\theta}}(\cdot)$.

Essentially, $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ summarizes the operation in each step with $\boldsymbol{\vartheta}$ being the input and $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ being the output when the sample size goes to infinity. By part (ii) of Lemma 1.4.1, $\overline{\boldsymbol{\theta}}(\cdot)$ is well-defined. Part (v) of Lemma 1.4.1 plays the most important role. Heuristically, for any given $\boldsymbol{\vartheta}$, the value of function $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ is obtained by solving $\frac{\partial Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$. At the same time, function $\frac{\partial Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}$ relates to the first order derivative of $L_{C|V,M}(\boldsymbol{\theta})$, the population objective function defined in Section 1.2.1. By the identification assumption and the convexity of $-L_{C|V,M}(\boldsymbol{\theta})$, only the true value $\boldsymbol{\theta}^*$ satisfies that $\frac{\partial}{\partial \boldsymbol{\theta}} L_{C|V,M}(\boldsymbol{\theta}) = \mathbf{0}$, which implies that $\frac{\partial Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \mathbf{0}$ holds only at $(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = (\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$.

If a consistent initial estimator is used, such as $\check{\boldsymbol{\theta}}$, then Lemma 1.4.1 is sufficient for the consistency of the IDC estimator as stated below.

Theorem 1.4.1 (Consistent initial value)**.** *Suppose Assumptions 1.4.1 and 1.4.2 hold. If* $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$, *then* $\widehat{\boldsymbol{\theta}}^{(S)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$ *for any* S.

On the other hand, if the initial estimator is consistent only under extra assumptions, such as $\widehat{\boldsymbol{\theta}}_T$ and $\widehat{\boldsymbol{\theta}}_P$, or even inconsistent, then we need a contraction mapping assumption on $\overline{\boldsymbol{\theta}}(\cdot)$.

Assumption 1.4.3 (Contraction Mapping)**.** *For any* $\boldsymbol{\vartheta} \in \Theta$, *there exists a constant* $C < 1$ *such that*

$$\left\|\overline{\boldsymbol{\theta}}\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \leq C \left\|\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) - \boldsymbol{\vartheta}\right\|.$$

Admittedly, Assumption 1.4.3 is a high-level assumption. Based on the evidence from the simulation, the assumption holds for various values of $\boldsymbol{\theta}^*$ and distributions of $\boldsymbol{V}$ and M. Assumption 1.4.3 relates to the contraction mapping assumption (Assumption 6) in [Pastorello et al., 2003] but is weaker, the reason being that the true $\boldsymbol{\theta}^*$ is the unique fixed point, see Lemma 1.4.1 (v). Specifically, Assumption 1.4.3 only requires that the distance

between $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ and $\boldsymbol{\vartheta}$ get smaller after both being mapped by $\overline{\boldsymbol{\theta}}(\cdot)$. Instead, the contraction mapping assumption (Assumption 6) in [Pastorello et al., 2003] requires that the distance between two arbitrary $\boldsymbol{\vartheta}^1$ and $\boldsymbol{\vartheta}^2$ get smaller after both being mapped by $\overline{\boldsymbol{\theta}}(\cdot)$.

Theorem 1.4.2 (Inconsistent initial value)**.** *Under Assumptions 1.4.1-1.4.3,* $\widehat{\boldsymbol{\theta}}^{(S)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$ *if* $S \to \infty$.

1.4.2 Asymptotic Distributions and Inference

Under Assumptions 1.4.1 and 1.4.2, the MLE $\widetilde{\boldsymbol{\theta}}$ is asymptotically normally distributed with asymptotic variance given by the Fisher information matrix

$$\mathcal{I}(\boldsymbol{\theta}^*) \equiv \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} Q^*(\boldsymbol{\theta}^*),$$

where $Q^*(\boldsymbol{\theta}^*) \equiv p\lim_{n\to\infty} \frac{1}{n} Q_n^*(\boldsymbol{\theta}^*)$. In this section, we show that our IDC estimator has the same asymptotic distribution as $\widetilde{\boldsymbol{\theta}}$, based on which we introduce a valid bootstrap inference procedure.

The conditions required for proving the asymptotic distribution result depend on the initial estimator $\widehat{\boldsymbol{\theta}}^{(0)}$. If a consistent initial estimator is used, then the following assumption is sufficient. For any matrix A, denote $\|A\|$ as its spectral norm.

Assumption 1.4.4 (Information Dominance)**.** *It holds that*

$$\left\| \left(\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*} \right)^{-1} \frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*} \right\| < 1.$$

The detailed expressions of $\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ and $\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'}$ can be found in Appendix A.1. Assumption 1.4.4 is often called the information dominance condition and is tantamount to the local contraction mapping condition. It is weaker than Assumption 1.4.3. Because we have an initial consistent estimator of $\boldsymbol{\theta}^*$, Assumption 1.4.4 can be verified. Additionally, because the matrix $\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ is block diagonal with each block having dimensions $p \times p$, computing

its inverse is feasible.

The following theorem shows that when S is sufficiently large, the IDC estimator $\widehat{\boldsymbol{\theta}}^{(S)}$ is equal to $\widetilde{\boldsymbol{\theta}}$ up to a term of order smaller than $n^{-1/2}$.

Theorem 1.4.3. *(i) Suppose Assumptions 1.4.1, 1.4.2, and 1.4.4 hold. If* $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$, *then* $\widehat{\boldsymbol{\theta}}^{(S)} - \widetilde{\boldsymbol{\theta}} = o_p\left(n^{-1/2}\right)$ *if* $S \geq \log(n)$. *(ii) Under Assumptions 1.4.1-1.4.3,* $\widehat{\boldsymbol{\theta}}^{(S)} - \widetilde{\boldsymbol{\theta}} = o_p\left(n^{-1/2}\right)$ *if* $S > n^{\delta}$ *for some* $\delta > 0$.

Theorem 1.4.3 shows that we do not lose efficiency when employing the proposed IDC estimator as long as S is large enough. A direct implication of the theorem is that $\widehat{\boldsymbol{\theta}}^{(S)}$ has the same asymptotic distribution as $\widetilde{\boldsymbol{\theta}}$ for sufficiently large S.

Corollary 1.4.4. *(i) Suppose Assumptions 1.4.1, 1.4.2, and 1.4.4 hold. If* $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$ *and* $S \geq \log(n)$, *then* $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}^*)\right)$ *as* $n \to \infty$. *(ii) Under Assumptions 1.4.1-1.4.3, if* $S > n^{\delta}$ *for some* $\delta > 0$, *then* $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}^*)\right)$ *as* $n \to \infty$.

To conduct inference on $\boldsymbol{\theta}^*$ based upon $\widehat{\boldsymbol{\theta}}^{(S)}$, we need to consistently estimate the Fisher information matrix and compute its inverse. Because the dimension of $\mathcal{I}(\boldsymbol{\theta}^*)$ is $dp \times dp$, calculating the inverse of its estimator is not only time-consuming but also unreliable when d is large. As a result, we proceed by applying the following parametric bootstrap, which is feasible thanks to the fact that the IDC estimator is fast to compute.

Given $\{(\boldsymbol{V}_i, M_i)\}_{i=1}^n$ of the original sample, we draw the bootstrap sample $\boldsymbol{C}_{in}^{\ddagger}$ for $i = 1, \ldots, n$ from the multinomial logistic regression model with the conditional probability mass function given by

$$\text{MNL}\left(\boldsymbol{C}_{in}^{\ddagger}; \widehat{\boldsymbol{\eta}}_i, M_i\right), \text{ where } \widehat{\eta}_{ik} = \boldsymbol{V}_i' \widehat{\boldsymbol{\theta}}_k^{(S)} \text{ for } k = 1, \ldots d.$$

The bootstrap version of the iterative estimator $\widehat{\boldsymbol{\theta}}^{\ddagger(S)}$ is obtained by applying the algorithm introduced in Section 1.3.2 with bootstrap sample $\left\{\left(\boldsymbol{C}_{in}^{\ddagger}, \boldsymbol{V}_i, M_i\right)\right\}_{i=1}^n$.

Assume that we start with a consistent initial estimator. Based on Theorem 1.4.3 (i), we have that for $S \geq \log(n)$,

$$\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\theta}^* = \mathcal{I}^{-1}(\boldsymbol{\theta}^*) \frac{1}{n} \frac{d}{d\boldsymbol{\theta}} l_{C|V,M}(\boldsymbol{\theta}^*) + o_p\left(n^{-1/2}\right). \tag{1.4.2}$$

Define the score function for $\boldsymbol{\theta}$ as

$$\dot{\mathbf{l}}(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{v}, m) \equiv \frac{d}{d\boldsymbol{\theta}} \log \mathrm{MNL}(\boldsymbol{c}; \boldsymbol{\eta}, m), \text{ where } \eta_k \equiv \boldsymbol{v}'\boldsymbol{\theta}_k.$$

The iterative estimator $\widehat{\boldsymbol{\theta}}^{(S)}$ is asymptotically linear with influence function $\widetilde{\mathbf{l}}(\boldsymbol{\theta}^* \mid \boldsymbol{c}, \boldsymbol{v}, m)$:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\theta}^*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\mathbf{l}}(\boldsymbol{\theta}^* \mid \boldsymbol{C}_i, \boldsymbol{V}_i, M_i) + o_p(1),$$

where $\widetilde{\mathbf{l}}(\boldsymbol{\theta}^* \mid \boldsymbol{c}, \boldsymbol{v}, m) \equiv \mathcal{I}^{-1}(\boldsymbol{\theta}^*) \dot{\mathbf{l}}(\boldsymbol{\theta}^* \mid \boldsymbol{c}, \boldsymbol{v}, m)$. Applying the same derivation, we can show that the bootstrap version of the estimator is also asymptotically linear with the influence function evaluated at $\widehat{\boldsymbol{\theta}}^{(S)}$:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{\ddagger(S)} - \widehat{\boldsymbol{\theta}}^{(S)}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\mathbf{l}}\left(\widehat{\boldsymbol{\theta}}^{(S)} \mid \boldsymbol{C}_{in}^{\ddagger}, \boldsymbol{V}_i, M_i\right) + o_p(1).$$

The Lindeberg-Feller central limit theorem proves the bootstrap consistency. The proof for the case of inconsistent initial estimator is analogous. Let $\xrightarrow{d^\ddagger}$ denote the convergence in bootstrap distribution.

Theorem 1.4.5. *(i) Suppose Assumptions 1.4.1, 1.4.2, and 1.4.4 hold. If* $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ *as* $n \to \infty$ *and* $S \geq \log(n)$, *then* $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{\ddagger(S)} - \widehat{\boldsymbol{\theta}}^{(S)}\right) \xrightarrow{d^\ddagger} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}^*)\right)$ *as* $n \to \infty$. *(ii) Under Assumptions 1.4.1-1.4.3, if* $S > n^\delta$ *for some* $\delta > 0$, *then* $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{\ddagger(S)} - \widehat{\boldsymbol{\theta}}^{(S)}\right) \xrightarrow{d^\ddagger} \mathcal{N}\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}^*)\right)$ *as* $n \to \infty$.

1.5 Monte Carlo Simulation

In this section, we evaluate the performance of our IDC estimator from various perspectives. We present the finite sample performance of the IDC estimator by looking at the effects of separately increasing d and n on mean squared error (MSE) and running time. We include the maximum likelihood estimator to show that the IDC estimator performs similarly to the MLE in terms of MSE and is always feasible even in cases where MLE is intractable. Lastly, we study the finite sample size and power of our bootstrap inference procedure.

1.5.1 Estimation

In what follows, we present results on the finite sample performance of the IDC estimator in terms of MSE and running time in four tables. The reported running times in the tables are obtained from a cluster of 25 AWS EC2 instances with 12 vCPUs and 16GB memory. Such a configuration can be formed on commonly used cloud computing platforms within minutes. We employ this basic configuration to illustrate that our IDC estimator achieves superior performance compared to existing estimators, even when computational resources are suboptimal for distributed computing. All of the results presented in this section are repeated five hundred times and averaged.

We first study the finite sample performance of the IDC estimator with consistent $\check{\boldsymbol{\theta}}$ initialization, specifically what happens to MSE (n increasing, d fixed) and running time (d increasing, n fixed). We consider the following data generating process (DGP):

DGP-A [MNL]: We set $p = 5$. The covariate vector $\boldsymbol{V}$ follows the standard normal distribution; M follows the discrete uniform distribution on $[20, 30]$; and the values of $\boldsymbol{\theta}^*$ are obtained by random draws from the standard normal distribution.

			$n = 500$		$n = 1000$		$n = 2000$	
$\widehat{\boldsymbol{\theta}}^{(0)}$	S	d	MSE	Time	MSE	Time	MSE	Time
$\check{\boldsymbol{\theta}}$	10	10	.0030	45s	.0020	97s	.0012	252s
		20	.0083	52s	.0052	113s	.0020	277s
		50	.0373	85s	.0182	174s	.0094	453s
		100	.0793	157s	.0381	349s	.0171	756s
		150	.1749	211s	.0585	455s	.0223	1007s
	40	10	.0037	168s	.0021	320s	.0012	672s
		20	.0081	196s	.0040	352s	.0019	739s
		50	.0365	320s	.0185	576s	.0091	1142s
		100	.0661	590s	.0336	890s	.0158	1948s
		150	.1815	770s	.0577	1184s	.0211	2822s

Table I: Finite sample performance of IDC estimator with $\check{\boldsymbol{\theta}}$ initialization

From Table I, we observe that the MSE of the IDC estimator decreases as the sample size increases. When the number of iterations S increases, we see an improvement in the MSE. However, the improvement is marginal, suggesting that the IDC estimator with $\check{\boldsymbol{\theta}}$ initialization stabilizes with only a few iterations. We also see from Table I that the running time is approximately a linear function of d. When the number of cores available does not exceed the number of choices, the additional computational cost of increasing d is very small. After the cores are fully occupied by the number of processes, the running time becomes approximately linear. The nominal value of running time depends on the hardware specifications.

	$n = 500$		$n = 1000$		$n = 2000$	
d	MSE	Time	MSE	Time	MSE	Time
10	.0040	36s	.0020	105s	.0015	350s
20	.0101	54s	.0045	152s	.0026	652s
50	.0263	96s	.0187	308s	.0098	1523s
100	.0852	250s	.0405	862s	.0151	4375s
150	.1179	457s	.0536	1523s	.0291	9352s

Table II: MSE and running time of MLE $\widetilde{\boldsymbol{\theta}}$

To compare the performance of our IDC estimator with the MLE $\widetilde{\boldsymbol{\theta}}$, we simulate the MSE and running time of $\widetilde{\boldsymbol{\theta}}$ from the same DGP and present the result in Table II.[7] It can be seen that the MSE of the IDC estimator with $\check{\boldsymbol{\theta}}$ initialization is very close to that of the MLE even when the number of iterations is only 10. Note that the main advantage of the parallel estimator is best observed for high enough d because for low d, the communication between parallel processes is unnecessary and hence parallel computing increases the running time unnecessarily. The superior performance of the IDC estimator is apparent when d is large. For instance, when $d = 150$ and $n = 2000$, the IDC estimator with $S = 10$ achieves a similar MSE as the MLE with only about one-tenth of the running time. For higher-dimensional cases, such as when d exceeds 150, computing the MLE becomes computationally intensive and may not be practical for many applications. In comparison, the IDC estimator with $S = 10$ demonstrates more efficient computation times, requiring approximately 5, 10, and 20 minutes for sample sizes $n = 500$, 1000, and 2000, respectively. Moreover, the running time for the IDC estimator can be further decreased if more compute instances are used. For example, using 96 instances, the running time of the IDC estimator for $d = 150$ can be further reduced to 34, 51, and 188 seconds for $n = 500$, 1000, 2000 respectively even for $S = 40$. Compared to the corresponding running time of MLE, the running time of the IDC

[7] We also write code to try estimators in [Böhning and Lindsay, 1988], [Böhning, 1992], and [Simon et al., 2013] for comparison. However, our simulation result suggests that their performance depends crucially on the number of iterations.

estimator using 96 instances is more than 10, 30, and 50 times shorter.

			$n = 500$		$n = 1000$		$n = 2000$	
$\widehat{\boldsymbol{\theta}}^{(0)}$	S	d	MSE	Time	MSE	Time	MSE	Time
$\widehat{\boldsymbol{\theta}}_T$	10	10	.0048	42s	.0023	95s	.0017	247s
		20	.0098	49s	.0058	106s	.0038	278s
		50	.0466	83s	.0197	175s	.0100	442s
		100	.0809	151s	.0381	339s	.0184	754s
		150	.1725	206s	.0590	451s	.0242	998s
	40	10	.0051	160s	.0025	311s	.0017	667s
		20	.0102	187s	.0052	337s	.0022	728s
		50	.0464	309s	.0198	564s	.0094	1128s
		100	.0867	581s	.0407	876s	.0179	1938s
		150	.1808	758s	.0588	1177s	.0219	2801s

			$n = 500$		$n = 1000$		$n = 2000$	
$\widehat{\boldsymbol{\theta}}^{(0)}$	S	d	MSE	Time	MSE	Time	MSE	Time
$\widehat{\boldsymbol{\theta}}_P$	10	10	.0068	52s	.0023	97s	.0016	244s
		20	.0118	59s	.0058	111s	.0038	290s
		50	.0466	83s	.0197	177s	.0100	438s
		100	.0808	155s	.0382	341s	.0184	751s
		150	.1739	209s	.0588	457s	.0258	1008s
	40	10	.0050	160s	.0025	322s	.0017	651s
		20	.0102	198s	.0052	355s	.0022	728s
		50	.0464	318s	.0198	570s	.0095	1140s
		100	.0866	577s	.0407	881s	.0179	1957s
		150	.1926	761s	.0587	1151s	.0245	2811s

Table III: Finite sample performance of IDC estimator with $\widehat{\boldsymbol{\theta}}_T$ and $\widehat{\boldsymbol{\theta}}_P$ initialization

In Table III, we present the MSE and running time of IDC estimators with $\widehat{\boldsymbol{\theta}}_T$ and $\widehat{\boldsymbol{\theta}}_P$ as initial estimators, respectively, for two different numbers of iterations S. Comparing MSEs of three IDC estimators with different initial values: $\check{\boldsymbol{\theta}}$ (Table I) and $\widehat{\boldsymbol{\theta}}_T$ (Table III) or $\widehat{\boldsymbol{\theta}}_P$ (Table III), we observe that the IDC estimator with the consistent initial estimator reduces

the MSEs for the same number of iterations.

	d	n	$\widetilde{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}^I_{PB}$	$\widehat{\boldsymbol{\theta}}^I_{P}$	$\widehat{\boldsymbol{\theta}}^I_{T}$
DGP-A	20	500	.0102	.0083	.0102	.0102
	20	1000	.0049	.0048	.0052	.0052
	50	1000	.0155	.0158	.0198	.0198
DGP-B	20	500	.0007	.0008	.0006	.0303
	20	1000	.0002	.0002	.0003	.0403
	50	1000	.0005	.0005	.0006	.0121
DGP-C	20	500	.0060	.0062	.0072	.0061
	20	1000	.0032	.0032	.0032	.0045
	50	1000	.0315	.0318	.0305	.0323

Table IV: MSE comparison of competing estimators. Number of iterations $S = 20$.

Table IV presents MSEs of different estimators for three (d, n) pairs each. We set the largest d be 50 so that MLE can be computed in a reasonable time. $\widehat{\boldsymbol{\theta}}^I_{PB}$, $\widehat{\boldsymbol{\theta}}^I_{T}$, and $\widehat{\boldsymbol{\theta}}^I_{P}$ denote the IDC estimators with $\check{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\theta}}_T$, and $\widehat{\boldsymbol{\theta}}_P$ as the initial estimators respectively. Besides DGP-A, we consider two additional DGPs to study the performance of the IDC estimator under different data settings. In all DGPs, we let $p = 5$.

DGP-B [Poisson]: The random variable $\boldsymbol{V}$ follows a standard normal distribution; C_{ik} follows a Poisson distribution with mean $e^{\eta^*_{ik}}$; and M is obtained by summing up realizations of the Poisson draws for different choices.

DGP-C [Mixture]: We let $\boldsymbol{V}$ follow a mixture of Gaussian distributions with means 0 and 4 with standard deviations 1 for both distributions. M is also set to follow a mixture of Gaussian distributions with means 10 and 60 and rounded to the closest integer. The standard deviations are 1 and 5 respectively. We have made these modifications so that some choices are rarely selected and ensure the robustness of our estimator in those cases.

Based on the simulation result, our IDC algorithm is successfully executed for all DGPs and exhibits stability. In contrast, we encounter errors for DGP-C when computing the

estimators in [Böhning and Lindsay, 1988], [Böhning, 1992], and [Simon et al., 2013]. We see from Table IV that $\widehat{\boldsymbol{\theta}}^I_{PB}$ performs close to $\widetilde{\boldsymbol{\theta}}$ for all DGPs and (d, n) pairs. In DGP-B, the initial estimator $\widehat{\boldsymbol{\theta}}_P$ is the maximum likelihood estimator. As a result, $\widehat{\boldsymbol{\theta}}^I_P$ starts with not only a consistent but asymptotically efficient initial estimator. Even in this case, $\widehat{\boldsymbol{\theta}}^I_{PB}$ has comparable MSEs.

In summary, the IDC estimators with all three initial estimators have finite sample performance similar to the MLE for the DGPs studied in this section. They are much faster to compute than the MLE for large d and are feasible even when the MLE might be intractable. Moreover, if the IDC estimator starts with the consistent initial estimator $\check{\boldsymbol{\theta}}$, its finite sample performance will be further improved and is almost the same as the MLE.

1.5.2 Inference

In this section, we illustrate the bootstrap inference procedure introduced in Section 1.4.2. We investigate the finite sample performance of the procedure including the size and power. All the results are based on one thousand Monte Carlo repetitions, where the number of bootstrap repetitions is five hundred.

We consider the null hypothesis that some element of $\boldsymbol{\theta}^*$ equals to a specific value. Data are generated from DGP-A introduced in the previous section. Let the null and the alternative hypotheses be that $H_0 : \theta^*_{11} = 0$ and $H_1 : \theta^*_{11} \neq 0$. The test statistic is computed as $\left|\frac{\widehat{\theta}^I_{11}}{\widehat{se}_b\left(\widehat{\theta}^I_{11}\right)/\sqrt{n}}\right|$, where $\widehat{se}_b\left(\widehat{\theta}^I_{11}\right)$ is the bootstrap estimate of the standard error of $\widehat{\theta}^I_{11}$. The number of iterations is 10 when computing the IDC estimator. We set the nominal size as 5% and use the 97.5% quantile of the standard normal distribution as the critical value.

	Dev.	-0.2	-0.1	-0.05	0	0.05	0.1	0.2
$d = 20$	$n = 250$	.397	.137	.067	.041	.060	.139	.455
	$n = 500$	.676	.216	.102	.058	.115	.234	.704
	$n = 1000$	.952	.395	.128	.055	.167	.473	.963
	Dev.	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
$d = 50$	$n = 250$	.333	.189	.099	.077	.091	.186	.313
	$n = 500$	.662	.424	.172	.066	.198	.391	.658
	$n = 1000$	.895	.723	.247	.063	.212	.697	.924

Table V: Finite sample rejection probabilities for different values of θ^*_{11}, n, and d

In Table V, we report the finite sample rejection probabilities of our test for different values of θ^*_{11}. Values in the first row of the table indicate the deviation of θ^*_{11} from the null hypothesis. When the deviation is zero, the null hypothesis is true. It can be seen from the table that the finite sample rejection rates get closer to the nominal size when the sample size increases. And when the true value θ^*_{11} deviates more from the null hypothesis, the rejection probabilities increase. The same pattern appears for both $d = 20$ and $d = 50$. The finite sample performance of the test when $d = 50$ is not as good as that when $d = 20$. This is predictable because there are many more unknown parameters in the model when $d = 50$ than when $d = 20$. But we expect the results to improve as the sample size increases for any fixed d.

1.6 Conclusion

In this paper, we propose an iterative distributed computing estimator for the multinomial logistic model that is fast to compute even when the number of choices is large. When the number of iterations goes to infinity, we show that our estimator is both consistent and asymptotically efficient. Based on the simulation study, the computational time of our estimator increases linearly with the number of choices. Moreover, our estimator has comparable finite sample performance to MLE when the latter is computationally feasible.

Extensions abound. First, our IDC estimator can be combined with several existing algorithms to accommodate more complex settings. For example, when minimizing $Q_{kn}\left(\boldsymbol{\theta}_k, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)$ for $k = 1, \ldots, d$, we can replace the gradient of the objective function with its stochastic approximation calculated from a randomly selected subset of the data. Such an algorithm is an online algorithm and might reduce the running time especially when n is large. We can also employ a one-step Newton-Raphson to compute $\arg\min_{\boldsymbol{\theta}_k \in \Theta_k} Q_{kn}\left(\boldsymbol{\theta}_k, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)$ in each iteration or even replace the Hessian matrix with its dominant. Depending on applications, these modifications of the IDC estimator may further improve the computational time. However, their theoretical properties need to be investigated. Second, asymptotic properties of the IDC estimator in this paper are established for large but fixed number of choices. Asymptotic theory allowing for the number of choices to diverge with the sample size is yet to be established. Third, in cases where the number of covariates is also large, ℓ_1 or ℓ_2 regularization could be adopted. In a companion paper, we develop asymptotic theory for regularized iterative distributed computing estimator.

.1 Notations and Equalities

In this appendix, we list some mathematical expressions and equalities used in the paper.

1. $\text{MNL}\left(\boldsymbol{C}_i;\boldsymbol{\eta}_i,M_i\right)\equiv\frac{M_i!}{C_{i1}!\cdots C_{id}!}\left(\frac{e^{\eta_{i1}}}{\Lambda_i}\right)^{C_{i1}}\ldots\left(\frac{e^{\eta_{id}}}{\Lambda_i}\right)^{C_{id}}$
2. $l_{C|V,M}\left(\boldsymbol{\theta}\right)\equiv\sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right]$
3. $L_{C|V,M}\left(\boldsymbol{\theta}\right)\equiv p\lim_{n\to\infty}\frac{1}{n}l_{C|V,M}\left(\boldsymbol{\theta}\right)$
4. $Q_n^*\left(\boldsymbol{\theta}\right)\equiv-l_{C|V,M}\left(\boldsymbol{\theta}\right)$
5. $\widetilde{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}\in\Theta}Q_n^*\left(\boldsymbol{\theta}\right)$
6. $l_{C|V,M}\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)\equiv\sum_{i=1}^{n}\left[\boldsymbol{C}_i'\left(\boldsymbol{\eta}_i+\mu_i\mathbf{1}_d\right)-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}+\mu_i}\right)\right]=l_{C|V,M}\left(\boldsymbol{\theta}\right)$
7. $f\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)\equiv\sum_{i=1}^{n}\left[M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)+M_i\mu_i-e^{\mu_i}\sum_{k=1}^{d}e^{\eta_{ik}}\right]$
8. $ql_{C|V}\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)\equiv l_{C|V,M}\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)+f\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)=\sum_{i=1}^{n}\sum_{k=1}^{d}\left(C_{ik}\left(\eta_{ik}+\mu_i\right)-e^{(\eta_{ik}+\mu_i)}\right)$
9. $Q_n\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)\equiv-ql_{C|V}\left(\boldsymbol{\theta},\boldsymbol{\mu}\right)$
10. $\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\mu}}\right)=\arg\min_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\mu}\in\mathbb{R}^n}Q_n\left(\boldsymbol{\theta},\boldsymbol{\mu}\right),$
11. $Q_{kn}\left(\boldsymbol{\theta}_k,\boldsymbol{\mu}\right)\equiv\sum_{i=1}^{n}\left(e^{\left(\boldsymbol{V}_i'\boldsymbol{\theta}_k+\mu_i\right)}-C_{ik}\left(\boldsymbol{V}_i'\boldsymbol{\theta}_k+\mu_i\right)\right)$
12. $\overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\theta}\right)\equiv\left(\log\left(\frac{M_1}{\sum_{k=1}^{d}e^{\eta_{1k}}}\right),\ldots,\log\left(\frac{M_n}{\sum_{k=1}^{d}e^{\eta_{nk}}}\right)\right)'$
13. $\widehat{\boldsymbol{\theta}}^{(s)}=\left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1}Q_{1n}\left(\boldsymbol{\theta}_1,\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right),\ldots,\arg\min_{\boldsymbol{\theta}_d\in\Theta_d}Q_{dn}\left(\boldsymbol{\theta}_d,\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)\right]'$
14. $l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right)\equiv\sum_{i=1}^{n}\left[C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k-\left(C_{ik}+C_{id}\right)\log\left(e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}+1\right)\right]$
15. $\check{\boldsymbol{\theta}}_k=\arg\min_{\boldsymbol{\theta}_k\in\Theta_k}-l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right)$

16. $\widehat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\mu}}_T\right)$ with $\widehat{\boldsymbol{\mu}}_T = \left(\log\left(M_1\right), \ldots, \log\left(M_n\right)\right)'$

17. $\widehat{\boldsymbol{\theta}}_P = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \mathbf{0}\right)$

18. $Q_n^{\dagger}\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right) \equiv Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\vartheta}\right)\right) = \sum_{i=1}^{n}\sum_{k=1}^{d}\left(\frac{M_i e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}_i'\boldsymbol{\vartheta}_k}} - C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k - C_{ik}\log\left(\frac{M_i}{\sum_{k=1}^{d} e^{\boldsymbol{V}_i'\boldsymbol{\vartheta}_k}}\right)\right)$

19. $Q^{\dagger}\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right) \equiv \sum_{k=1}^{d}\mathbb{E}\left[\left(\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k - C_k\log\left(\frac{M}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\right)\right)\right]$

20. $Q_k^{\dagger}\left(\boldsymbol{\theta}_k, \boldsymbol{\vartheta}\right) \equiv \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k - C_k\log\left(\frac{M}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\right)\right]$

21. $\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}\right) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\dagger}\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right)$

22. $Q^{*}\left(\boldsymbol{\theta}^{*}\right) \equiv p\lim_{n\to\infty}\frac{1}{n}Q_n^{*}\left(\boldsymbol{\theta}^{*}\right)$

23. $\mathcal{I}\left(\boldsymbol{\theta}^{*}\right) \equiv \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q^{*}\left(\boldsymbol{\theta}^{*}\right)$

24. $\frac{\partial^2 Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \equiv \begin{bmatrix} \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_d}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] \end{bmatrix}$

25. $\frac{\partial^2 Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'} \equiv \begin{bmatrix} \mathbb{E}\left[-\frac{Me^{2\boldsymbol{V}'\boldsymbol{\theta}_1}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbb{E}\left[-\frac{Me^{\boldsymbol{V}'(\boldsymbol{\theta}_1+\boldsymbol{\theta}_d)}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[-\frac{Me^{\boldsymbol{V}'(\boldsymbol{\theta}_1+\boldsymbol{\theta}_d)}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbb{E}\left[\frac{Me^{2\boldsymbol{V}'\boldsymbol{\theta}_d}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] \end{bmatrix}$

26. $\dot{\mathrm{I}}\left(\boldsymbol{\theta} \mid \boldsymbol{C}_i, \boldsymbol{V}_i, M_i\right) \equiv \frac{d}{d\boldsymbol{\theta}}\log\mathrm{MNL}\left(\boldsymbol{C}_i; \boldsymbol{\eta}_i, M_i\right)$

27. $\widetilde{\mathrm{I}}\left(\boldsymbol{\theta} \mid \boldsymbol{C}_i, \boldsymbol{V}_i, M_i\right) \equiv \mathcal{I}^{-1}\left(\boldsymbol{\theta}\right)\dot{\mathrm{I}}\left(\boldsymbol{\theta} \mid \boldsymbol{C}_i, \boldsymbol{V}_i, M_i\right)$

.2 Proofs

Proof of Lemma 1.2.1: It holds that

$$\begin{aligned} l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu}) &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i+\mu_i\boldsymbol{C}_i'\mathbf{1}_n-M_i\log\left(e^{\mu_i}\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\ &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i+\mu_i\sum_{k=1}^{d}C_{ik}-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)-M_i\mu_i\right] \\ &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\ &= l_{C|V,M}(\boldsymbol{\theta}), \end{aligned}$$

where the second to last equality holds because $\mu_i\sum_{k=1}^{d}C_{ik}=M_i\mu_i$. Therefore, adding μ_i does not change the likelihood. □

Proof of Lemma 1.2.2: Notice that $Q_n(\boldsymbol{\theta},\boldsymbol{\mu})$ is differentiable w.r.t. $\boldsymbol{\mu}$ for any given $\boldsymbol{\theta}$. By letting $\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}=\mathbf{0}$, we can obtain function $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$ such that $\left.\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})}=\mathbf{0}$ for every $\boldsymbol{\theta}$. By definition,

$$Q_n(\boldsymbol{\theta},\boldsymbol{\mu})=-\left[l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu})+f(\boldsymbol{\theta},\boldsymbol{\mu})\right]=-\left[l_{C|V,M}(\boldsymbol{\theta})+f(\boldsymbol{\theta},\boldsymbol{\mu})\right],$$

which implies that $\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}=\frac{\partial f(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}$. Since $\frac{\partial f(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\mu_i}=M_i-e^{\mu_i}\sum_{k=1}^{d}e^{\eta_{ik}}$, we obtain the expression of $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$ as in (1.3.3). Plugging it into (1.2.5), we have that

$$\begin{aligned} & f(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})) \\ = & -\sum_{i=1}^{n}\left[M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)+M_i\log\left(\frac{m_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)-e^{\log\left(\frac{M_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)}\sum_{k=1}^{d}e^{\eta_{ik}}\right] \\ = & -\sum_{i=1}^{n}\left[M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)+M_i\log M_i-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)-\left(\frac{M_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)\sum_{k=1}^{d}e^{\eta_{ik}}\right] \\ = & -\sum_{i=1}^{n}\left[M_i\log M_i-M_i\right], \end{aligned} \tag{.2.1}$$

which does not depend on $\boldsymbol{\theta}$. As a result, it holds that

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\theta}\right)\right) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta}\left[Q_n^*\left(\boldsymbol{\theta}\right) - f\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\theta}\right)\right)\right] = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*\left(\boldsymbol{\theta}\right).$$

Because $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}\left(\boldsymbol{\theta}\right)\right)$, we have that

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*\left(\boldsymbol{\theta}\right) = \widetilde{\boldsymbol{\theta}}.$$

The claimed lemma then follows. □

Proof of Lemma 1.3.1: By Equation (1.2.1), it holds that

$$\begin{aligned}
&\Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, M_i\right) \\
=&\frac{M_i!}{C_{ik}!C_{id}!\left(M_i - C_{ik} - C_{id}\right)!}\left(\frac{e^{\eta_{ik}^*}}{\Lambda_i^*}\right)^{C_{ik}}\left(\frac{e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{C_{id}}\left(\frac{\Lambda_i^* - e^{\eta_{ik}^*} - e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{M_i - C_{ik} - C_{id}}. \qquad (.2.2)
\end{aligned}$$

Because $N_{ik} = C_{ik} + C_{id}$, we have that

$$\begin{aligned}
\Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, M_i\right) &= \Pr\left(C_{ik}, C_{id}, N_{ik} \mid \boldsymbol{V}_i, M_i\right) \\
&= \Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i, M_i\right)\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right). \qquad (.2.3)
\end{aligned}$$

Compute $\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right)$ as

$$\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right) = \frac{M_i!}{N_{ik}!\left(M_i - N_{ik}\right)!}\left(\frac{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{N_{ik}}\left(\frac{\Lambda_i^* - e^{\eta_{ik}^*} - e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{M_i - N_{ik}}.$$

Together with Equations (A.2.2) and (A.2.3), we obtain that

$$\begin{aligned}
\Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i, M_i\right) &= \frac{N_{ik}!}{C_{ik}!C_{id}!}\left(\frac{e^{\eta_{ik}^*}}{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}\right)^{C_{ik}}\left(\frac{e^{\eta_{id}^*}}{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}\right)^{C_{id}} \\
&= \Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i\right).
\end{aligned}$$

By replacing $e^{\eta_{id}^*}$ with 1 because $\boldsymbol{\theta}_d^* = \mathbf{0}$, we obtain the claimed result. □

Proof of Lemma 1.3.2: Under the assumption that $\Pr(M_i \mid \boldsymbol{V}_i) = \text{Po}\left(\sum_{k=1}^d e^{\eta_{ik}^*}\right)$, we can obtain that

$$\begin{aligned}\Pr(\boldsymbol{C}_i \mid \boldsymbol{V}_i) &= \Pr(\boldsymbol{C}_i \mid \boldsymbol{V}_i, M_i)\,\text{Po}\left(\sum_{k=1}^d e^{\eta_{ik}^*}\right) \\ &= \prod_{k=1}^d \text{Po}\left(e^{\eta_{ik}^*}\right) = \prod_{k=1}^d \frac{e^{\eta_{ik}^* C_{ik}} e^{-e^{\eta_{ik}^*}}}{C_{ik}!} = \frac{\prod_{k=1}^d e^{\eta_{ik}^* C_{ik}}}{C_{i1}! \cdots C_{id}!} e^{-\sum_{k=1}^d e^{\eta_{ik}^*}}.\end{aligned}$$

The log-likelihood function is written as

$$\begin{aligned}\log\left[\prod_{i=1}^n \frac{\prod_{k=1}^d e^{\eta_{ik} C_{ik}}}{C_{i1}! \cdots C_{id}!} e^{-\sum_{k=1}^d e^{\eta_{ik}}}\right] &\sim \sum_{i=1}^n \left[\sum_{k=1}^d C_{ik}\eta_{ik} - \sum_{k=1}^d e^{\eta_{ik}}\right] = \sum_{i=1}^n \sum_{k=1}^d \left(C_{ik}\eta_{ik} - e^{\eta_{ik}}\right) \\ &= ql_{C|V}(\boldsymbol{\theta}, \mathbf{0}).\end{aligned}$$

Therefore, $\widehat{\boldsymbol{\theta}}_P$ maximizes the true log-likelihood function based upon $\Pr(\boldsymbol{C}_i \mid \boldsymbol{V}_i)$. The lemma follows. □

Lemma .2.1. *Under Assumption 1.4.2, $-L_{C|V,M}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$.*

Proof of Lemma A.2.1: By the definitions of $l_{C|V,M}(\boldsymbol{\theta})$ and $L_{C|V,M}(\boldsymbol{\theta})$ and Assumption 1.4.2, we have that

$$\begin{aligned}
-\frac{1}{n} l_{C|V,M}(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{i=1}^{n}\left[C_{i1}\eta_{i1}+\cdots+C_{id}\eta_{id}-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[C_{i1}\boldsymbol{V}_i'\boldsymbol{\theta}_1+\cdots+C_{id}\boldsymbol{V}_i'\boldsymbol{\theta}_d-M_i\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}\right)\right] \\
&\xrightarrow{p} -\left(\mathbb{E}\left[C_1\boldsymbol{V}'\boldsymbol{\theta}_1\right]+\cdots+\mathbb{E}\left[C_d\boldsymbol{V}'\boldsymbol{\theta}_d\right]+\mathbb{E}\left[M\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)\right]\right) \\
&\equiv -L_{C|V,M}(\boldsymbol{\theta}) \\
&= -\sum_{k=1}^{d}\mathbb{E}\left[C_k\boldsymbol{V}'\boldsymbol{\theta}_k\right]+\mathbb{E}\left[M\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)\right].
\end{aligned}$$

The first term is convex in $\boldsymbol{\theta}$ because it only involves linear functions. It has been shown in Section 3.1.5 in [Boyd et al., 2004] that $\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)$ is convex in $\boldsymbol{V}'\boldsymbol{\theta}_1,\ldots,\boldsymbol{V}'\boldsymbol{\theta}_d$. Because $\boldsymbol{V}'\boldsymbol{\theta}_k$ is a linear function of θ_k and sums of convex functions are convex, the second term is convex in $\boldsymbol{\theta}$. □

Lemma .2.2. *$\boldsymbol{\theta}^*$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\boldsymbol{\theta})$.*

Proof of Lemma A.2.2: By definition, $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})$. Thus, if we can show that $\frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \mathbf{0}$ holds only at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, then $\boldsymbol{\theta}^*$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\cdot)$. Taking the first order derivative of $-L_{C|V,M}(\cdot)$, we obtain that for any $\dot{\boldsymbol{\theta}}\in\Theta$,

$$\begin{aligned}
-\frac{d}{d\boldsymbol{\theta}} L_{C|V,M}(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\dot{\boldsymbol{\theta}}} &= \left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\dot{\boldsymbol{\theta}}_1}\boldsymbol{V}'}{\sum_{k=1}^{d}e^{\boldsymbol{V}'\dot{\boldsymbol{\theta}}_k}}-C_1\boldsymbol{V}'\right],\ldots,\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\dot{\boldsymbol{\theta}}_{d-1}}\boldsymbol{V}'}{\sum_{k=1}^{d}e^{\boldsymbol{V}'\dot{\boldsymbol{\theta}}_k}}-C_{d-1}\boldsymbol{V}'\right]\right]' \\
&= \frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=(\dot{\boldsymbol{\theta}},\dot{\boldsymbol{\theta}})}.
\end{aligned}$$

Therefore, proving that $\frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \mathbf{0}$ holds only at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ is equivalent to showing that $\frac{d}{d\boldsymbol{\theta}} L_{C|V,M}(\boldsymbol{\theta}) = \mathbf{0}$ only at $\boldsymbol{\theta}^*$. By Lemma A.2.1, $-L_{C|V,M}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. Therefore, $-L_{C|V,M}(\boldsymbol{\theta})$ only has minimums in the interior of Θ. In addition, for a convex function over

a convex set, any local minimum is also a global minimum. Along with the identification assumption that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} -L_{C|V,M}(\boldsymbol{\theta})$, we obtain that $\frac{\partial}{\partial\boldsymbol{\theta}} L_{C|V,M}(\boldsymbol{\theta}) = \mathbf{0}$ holds only at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. □

Proof of Lemma 1.4.1: Part (i) follows by applying Theorem 2 in [Jennrich, 1969] on the uniform law of large numbers with conditions satisfied by Assumption 1.4.2.

Since $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is continuous in $\boldsymbol{\theta}$ and Θ is compact by Assumption 1.4.2 (i), $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ achieves its minimum in Θ for any $\boldsymbol{\vartheta}$. Given any $0 < \lambda < 1$ and $\boldsymbol{\theta}^1 \neq \boldsymbol{\theta}^2$, if $v_j \neq 0$ for $j = 1, \ldots, p$, then we have

$$e^{\boldsymbol{v}'\left[\lambda\boldsymbol{\theta}^1 + (1-\lambda)\boldsymbol{\theta}^2\right]} < \lambda e^{\boldsymbol{v}'\boldsymbol{\theta}^1} + (1-\lambda)\, e^{\boldsymbol{v}'\boldsymbol{\theta}^2}.$$

In consequence, $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta}$. Because Θ is a convex set, we have that $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ have a unique minimizer for any $\boldsymbol{\vartheta}$. Part (ii) holds.

For part (iii), because $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta} \in \Theta$ and Θ is convex, the (opposite) maximum theorem implies the continuity.

To prove part (iv), we take the first order partial derivative of $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ with respect to $\boldsymbol{\theta}$ and obtain that

$$\frac{\partial Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \left[\mathbb{E}\left[\frac{M e^{\boldsymbol{V}'\boldsymbol{\theta}_1}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_1\boldsymbol{V}'\right], \ldots, \mathbb{E}\left[\frac{M e^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_{d-1}\boldsymbol{V}'\right]\right]'.$$

Since $\mathbb{E}\left[C_k \mid \boldsymbol{V}, M\right] = \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}$, we have that

$$\begin{aligned}
&\frac{\partial Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right)}{\partial \boldsymbol{\theta}} \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)} \\
&= \left[\mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - C_1\boldsymbol{V}' \mid \boldsymbol{V}, M\right]\right], \ldots, \mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - C_d\boldsymbol{V}' \mid \boldsymbol{V}, M\right]\right]\right]' \\
&= \left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\right], \ldots, \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\right]\right]' \\
&= \mathbf{0}. \qquad (.2.4)
\end{aligned}$$

Because $Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right)$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta}$, $Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right)$ is certainly strictly convex in $\boldsymbol{\theta}$. Combining with (A.2.4), we obtain that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right)$, which implies that $\overline{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^*\right) = \boldsymbol{\theta}^*$.

Part (v) is proved by Lemma A.2.2. □

Proof of Theorem 1.4.2: For part (i), we show that if $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ as $n \to \infty$, then $\widehat{\boldsymbol{\theta}}^{(1)} \xrightarrow{p} \boldsymbol{\theta}^*$ as well. By (1.3.4), $\widehat{\boldsymbol{\theta}}_1^{(1)}$ satisfies that

$$\widehat{\boldsymbol{\theta}}_1^{(1)} = \arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right).$$

The first order condition provides that

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\widehat{\boldsymbol{\theta}}_1^{(1)}} = \mathbf{0}.$$

The mean value theorem implies that

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^*} + \frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^*} \left(\widehat{\boldsymbol{\theta}}_1^{(1)} - \boldsymbol{\theta}_1^*\right) \\
&\equiv A_n + B_b\left(\widehat{\boldsymbol{\theta}}_1^{(1)} - \boldsymbol{\theta}_1^*\right),
\end{aligned}$$

where $\boldsymbol{\theta}_1^\star$ lies between $\widehat{\boldsymbol{\theta}}_1^{(1)}$ and $\boldsymbol{\theta}_1^*$. Since $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$, we have that

$$A_n \xrightarrow{p} \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\boldsymbol{V} - C_k\boldsymbol{V}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\boldsymbol{V} - C_k\boldsymbol{V} \mid \boldsymbol{V}, M\right]\right] = \mathbf{0}.$$

By a similar argument, it can be shown that B_n converges in probability to a non-singular matrix for any $\boldsymbol{\theta}_1^\star \in \Theta_1$. Therefore, it must hold that $\widehat{\boldsymbol{\theta}}_1^{(1)} \xrightarrow{p} \boldsymbol{\theta}_1^*$. Hence, $\widehat{\boldsymbol{\theta}}^{(S)} \xrightarrow{p} \boldsymbol{\theta}^*$ as $n \to \infty$ for any S.

We now prove part (ii) of the theorem. Based on Lemma 1.4.1, Assumptions 1, 2a, and 5 in [Pastorello et al., 2003] are satisfied. Therefore, Proposition 1 in [Pastorello et al., 2003] holds, which implies that as $n \to \infty$,

$$\sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \xrightarrow{p} 0, \tag{.2.5}$$

where $\overline{\boldsymbol{\theta}}_n(\cdot)$ is defined in (1.3.4).

Let $\boldsymbol{\vartheta}^{(0)} \equiv p\lim_{n\to\infty} \widehat{\boldsymbol{\theta}}^{(0)}$, where $\widehat{\boldsymbol{\theta}}^{(0)}$ is the initial value of our IDC estimator $\widehat{\boldsymbol{\theta}}^I$. It can be seen that

$$\boldsymbol{\vartheta}^{(0)} = \left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_1(\boldsymbol{\theta}_1), \ldots, \arg\min_{\boldsymbol{\theta}_d\in\Theta_d} Q_d(\boldsymbol{\theta}_d)\right]',$$

where $Q_k(\boldsymbol{\theta}_k) \equiv \mathbb{E}\left[e^{\boldsymbol{V}'\boldsymbol{\theta}_k} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k\right]$. Define $\boldsymbol{\vartheta}^{(1)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(0)}\right)$, $\boldsymbol{\vartheta}^{(2)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(1)}\right) \equiv \overline{\boldsymbol{\theta}}^2\left(\boldsymbol{\vartheta}^{(0)}\right)$, ..., $\boldsymbol{\vartheta}^{(s)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(s-1)}\right) \equiv \overline{\boldsymbol{\theta}}^s\left(\boldsymbol{\vartheta}^{(0)}\right)$ for any $s \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ denotes the set of positive integers. Next, we show that $\left(\boldsymbol{\vartheta}^{(s)}\right)$ is a Cauchy sequence. By Assumption 1.4.3, we have that for any

$s_1 > s_2 \geq 1$,

$$\begin{aligned}
\left\|\boldsymbol{\vartheta}^{(s_1)} - \boldsymbol{\vartheta}^{(s_2)}\right\| &= \left\|\overline{\boldsymbol{\theta}}^{s_1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\| \\
&\leq \left[\left\|\overline{\boldsymbol{\theta}}^{s_1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_1-1}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\| + \left\|\overline{\boldsymbol{\theta}}^{s_1-1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_1-2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\|\right. \\
&\quad \left. + \cdots + \left\|\overline{\boldsymbol{\theta}}^{s_2+1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\|\right] \\
&\leq \left[C^{s_1-1} + C^{s_1-2} + \cdots + C^{s_2}\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&= C^{s_2}\left[\sum_{i=0}^{s_1-s_2-1} C^i\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&\leq C^{s_2}\left[\sum_{i=0}^{\infty} C^i\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&\leq \frac{C^{s_2}}{1-C}\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\|, \qquad (.2.6)
\end{aligned}$$

which implies that $\left(\boldsymbol{\vartheta}^{(s)}\right)$ is Cauchy because $C < 1$. Since $\Theta \subseteq \mathbb{R}^{p\times d}$ is compact by Assumption 1.4.2 and $\mathbb{R}^{p\times d}$ is complete with respect to $\|\cdot\|$, Θ is also complete with respect to $\|\cdot\|$. Therefore, $\boldsymbol{\vartheta}^{(s)}$ converges to a limit $\boldsymbol{\vartheta}^{\star}$ in Θ as $s \to \infty$. Because

$$\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{\star}\right) = \overline{\boldsymbol{\theta}}\left(\lim_{s\to\infty}\boldsymbol{\vartheta}^{(s)}\right) = \lim_{s\to\infty}\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(s)}\right) = \lim_{s\to\infty}\boldsymbol{\vartheta}^{(s+1)} = \boldsymbol{\vartheta}^{\star},$$

it holds that $\boldsymbol{\vartheta}^{\star}$ is a fixed point of the mapping $\overline{\boldsymbol{\theta}} : \Theta \to \Theta$. By Lemma A.2.2, $\boldsymbol{\theta}^{*}$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\cdot)$. Thus, $\boldsymbol{\vartheta}^{\star} = \boldsymbol{\theta}^{*}$ and $\lim_{s\to\infty}\boldsymbol{\vartheta}^{(s)} = \boldsymbol{\theta}^{*}$.

We now show that $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\vartheta}^{(s)} = o_p(1)$ for any $s \in \mathbb{Z}^{+}$ by induction. By the definition of $\boldsymbol{\vartheta}^{(0)}$, $\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\vartheta}^{(0)} = o_p(1)$. Assuming that $\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\vartheta}^{(t)} = o_p(1)$ for some t, it holds that

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\vartheta}^{(t+1)} &= \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(t)}\right) \\
&= \left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)\right] + \left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(t)}\right)\right] \\
&= o_p(1),
\end{aligned}$$

where the last equality holds because $\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) = o_p(1)$ by (A.2.5), $\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\vartheta}^{(t)} = o_p(1)$ by assumption and $\overline{\boldsymbol{\theta}}(\cdot)$ is continuous by Lemma 1.4.1 (iii). Therefore, $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\vartheta}^{(s)} = o_p(1)$ for any $s \in \mathbb{Z}^+$.

Hence, we have that if $S \to \infty$ and $n \to \infty$, then

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}^{I} - \boldsymbol{\theta}^*\right\| &= \left\|\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\vartheta}^{(S)}\right\| + \left\|\boldsymbol{\vartheta}^{(S)} - \boldsymbol{\theta}^*\right\| \\ &\equiv A(n,S) + B(S) \xrightarrow{p} 0,\end{aligned}$$

because $A(n,S) \xrightarrow{p} 0$ as $n \to \infty$ for any given S and $B(S) \to 0$ as $S \to \infty$ by $\lim_{s\to\infty} \boldsymbol{\vartheta}^{(s)} = \boldsymbol{\theta}^*$. The second part of the theorem holds. □

Lemma .2.3. *Under the conditions in Theorem 1.4.3, it holds that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right) \xrightarrow{p} 0.$$

Proof of Lemma A.2.3: We first show that the result in the lemma holds if the conditions in part (i) of the theorem hold. By (1.3.4), for any $s \in \mathbb{Z}^+$, $\widehat{\boldsymbol{\theta}}^{(s)}$ satisfies that

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}} = \frac{\partial}{\partial \boldsymbol{\theta}} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} = \mathbf{0}.$$

Apply Taylor expansion to the left-hand-side of the equality at $\boldsymbol{\theta}^*$. Because $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^* = o_p(1)$ and $\widehat{\boldsymbol{\theta}}^{(s-1)} - \boldsymbol{\theta}^* = o_p(1)$, we obtain that

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\theta}} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*, \boldsymbol{\vartheta}=\boldsymbol{\theta}^*} + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} \left(\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*\right) \\ &+ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\vartheta}'} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} \left(\widehat{\boldsymbol{\theta}}^{(s-1)} - \boldsymbol{\theta}^*\right) = \mathbf{0} \qquad (.2.7)\end{aligned}$$

by ignoring higher order terms. Since $\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}$ is non-singular and both $\widehat{\boldsymbol{\theta}}^{(s)}$ and $\widehat{\boldsymbol{\theta}}^{(s-1)}$ are consistent estimators of $\boldsymbol{\theta}^*$, we have that

$$\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}$$

exists with high probability when n is large. Define

$$\begin{aligned}
A_n &= \left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}\left(-\frac{\partial}{\partial\boldsymbol{\theta}}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)\\
B_n &= \left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}\left(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right).
\end{aligned}$$

By the law of large numbers and the consistency of $\widehat{\boldsymbol{\theta}}^{(s)}$ for any s, we have that

$$\begin{aligned}
A_n &= \left(\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)^{-1}\left(-\frac{\partial}{\partial\boldsymbol{\theta}}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)+o_p\left(1\right)\equiv A+o_p\left(1\right)\\
B_n &= \left(\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)^{-1}\left(-\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'}\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)+o_p\left(1\right)\equiv B+o_p\left(1\right).
\end{aligned}$$

By ignoring the smaller order terms, we obtain from Equation (A.2.7) that

$$\widehat{\boldsymbol{\theta}}^{(s)}-\boldsymbol{\theta}^*=A+B\left(\widehat{\boldsymbol{\theta}}^{(s-1)}-\boldsymbol{\theta}^*\right)=\sum_{t=0}^{s-1}B^tA+B^s\left(\widehat{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right),$$

where the second equality follows from iterating the first equality. It then holds that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}\right)=\sqrt{n}B^SA+\sqrt{n}B^S\left(B-I\right)\left(\widehat{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right).$$

By Assumption 1.4.4, $\|B\|<1$, which implies that $\sqrt{n}\left\|B\right\|^S\to 0$ as $S\geq\log\left(n\right)$ and $n\to\infty$. Hence, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}\right)=o_p\left(1\right)$. The claimed lemma follows.

We now prove that under the conditions in part (ii) of the theorem, the result also holds. If we can show that for any $s\in\mathbb{Z}^+$, with probability approaching one there exists a constant

$c < 1$ such that

$$\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \leq c\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|, \tag{.2.8}$$

then by the same derivation as (A.2.6), we would obtain that

$$\left\|\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right\| \leq \frac{c^S}{1-c}\left\|\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}^{(0)}\right\|.$$

Because $c < 1$ and $S > n^{\delta}$ for some $\delta > 0$, it holds that

$$\sqrt{n}\left\|\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right\| \leq \frac{n^{\frac{1}{2}}c^S}{1-c}\left\|\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}^{(0)}\right\| \to 0.$$

Thus, it suffices to prove (A.2.8).

By the implicit function theorem, $\overline{\boldsymbol{\theta}}(\cdot)$ is continuously differentiable in Θ. Together with Assumption 1.4.3, we have that there exists $\epsilon > 0$ such that for any $\boldsymbol{\vartheta} \in \Theta^d$ and $\check{\boldsymbol{\vartheta}} \in \mathcal{B}^{\epsilon}\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right) \equiv \left\{\boldsymbol{\theta} \in \Theta^d : \left\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \leq \epsilon\right\}$, we have

$$\left\|\overline{\boldsymbol{\theta}}(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \leq C_{\epsilon}\left\|\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\right\|,$$

for some $C_{\epsilon} \leq \overline{C} < 1$. By (A.2.5), $\Pr\left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right) \in \mathcal{B}^{\epsilon}\left(\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right)\right] \to 1$ as $n \to \infty$ for any s and ϵ. Therefore, with probability approaching one, it holds that

$$\begin{aligned}\left\|\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| &= \left\|\overline{\boldsymbol{\theta}}\left(\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \\ &\leq C_{\epsilon}\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right) - \widehat{\boldsymbol{\theta}}^{(s)}\right\| = C_{\epsilon}\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|.\end{aligned} \tag{.2.9}$$

For any $\boldsymbol{\vartheta} \in \Theta^d$, define $\Omega(\boldsymbol{\vartheta}) \equiv \frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$, where $\frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ is the Jacobian matrix of dimension $dp \times dp$. By the implicit function theorem, we have

$$\frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) = \left[\frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\right)}\right]^{-1} \frac{\partial}{\partial \boldsymbol{\vartheta}} g(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\right)},$$

where $g(\boldsymbol{\theta},\boldsymbol{\vartheta}) \equiv \frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}$ is the function that defines $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$. Similarly, we define $\Omega_n(\boldsymbol{\vartheta}) \equiv \frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta})$ and $g_n(\boldsymbol{\theta},\boldsymbol{\vartheta}) \equiv \frac{\partial Q_n^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}$. By Assumption 1.4.2 (ii) and (iii) and the uniform law of large numbers, we have

$$\sup_{\boldsymbol{\vartheta}\in\Theta} \|\Omega_n(\boldsymbol{\vartheta}) - \Omega(\boldsymbol{\vartheta})\| \xrightarrow{p} 0. \tag{.2.10}$$

By Assumption 1.4.2 (i), Θ is convex. Applying a multivariate Taylor expansion ([Dieudonné, 2011], p. 190), we can write $\overline{\boldsymbol{\theta}}(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) = \Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ for any $\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}} \in \Theta$, where $\Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) \equiv \int_0^1 \Omega(\boldsymbol{\vartheta} + \xi(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}))\, d\xi$. Similarly, we have $\overline{\boldsymbol{\theta}}_n(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta}) = \Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ for any $\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}} \in \Theta^d$, where $\Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) \equiv \int_0^1 \Omega_n(\boldsymbol{\vartheta} + \xi(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}))\, d\xi$. It holds that

$$\begin{aligned}
\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| &\leq \left\|\left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right] - \left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right]\right\| \\
&\quad + \left\|\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \\
&\leq \left\|\left[\Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) - \Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})\right]\left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right]\right\| \\
&\quad + C_{\epsilon}\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|,
\end{aligned}$$

where the first inequality follows from the triangular inequality and the second inequality holds by (A.2.9). Because the first term on the right hand side has the order $o_p\left(\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|\right)$ because of (A.2.10), we have shown that (A.2.8) holds with $c = C_{\epsilon} < 1$. The lemma follows. □

Proof of Theorem 1.4.3: We aim to show that $\widehat{\boldsymbol{\theta}}^{(S)}$ has the same influence function as $\widetilde{\boldsymbol{\theta}}$. By the definition of the IDC estimator in Section 1.3.2, we have that

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right) |_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S)}} = \mathbf{0}.$$

Applying the Taylor expansion to the function on the left-hand-side of the above equation round $\widehat{\boldsymbol{\theta}}^{(S-1)}$, we can obtain that

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}\\&+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}}\left(\widehat{\boldsymbol{\theta}}^{(S)}-\widehat{\boldsymbol{\theta}}^{(S-1)}\right)=\mathbf{0},\end{aligned}\tag{.2.11}$$

where $\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}$ lies between $\widehat{\boldsymbol{\theta}}^{(S)}$ and $\widehat{\boldsymbol{\theta}}^{(S-1)}$. The definition of $Q_n(\boldsymbol{\theta},\boldsymbol{\mu})$ implies that

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}\\&=-\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)-\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}.\end{aligned}$$

Applying the expression of $f(\boldsymbol{\theta},\boldsymbol{\mu})$, it can be shown that

$$\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}=\mathbf{0}.$$

Therefore, (A.2.11) can be rewritten as

$$-\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}}\left(\widehat{\boldsymbol{\theta}}^{(S)}-\widehat{\boldsymbol{\theta}}^{(S-1)}\right)=\mathbf{0}.$$

By Lemma A.2.3, we have $\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}=o_p\left(n^{-1/2}\right)$. This implies that

$$\begin{aligned}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)&=o_p\left(n^{-1/2}\right)\\&=\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^*\right)+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^\dagger\right)\left(\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^*\right),\end{aligned}\tag{.2.12}$$

where the second equality follows from the Taylor expansion and $\boldsymbol{\theta}^\dagger$ lies between $\widehat{\boldsymbol{\theta}}^{(S-1)}$ and $\boldsymbol{\theta}^*$.

By Theorem 1.4.2, we have that $\widehat{\boldsymbol{\theta}}^{(S-1)}\xrightarrow{p}\boldsymbol{\theta}^*$ as $n\to\infty$. Therefore, by Assumption 1.4.2

and Taylor's theorem, we can obtain that

$$\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{\dagger}\right)=\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(1\right).$$

Since matrix inversion is continuous (at non-singular matrices), it follows that the inverse of $\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)$ exists with high probability and

$$\left[-\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)\right]^{-1}\xrightarrow{p}\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right),$$

where $\mathcal{I}\left(\boldsymbol{\theta}^{*}\right)$ is the Fisher information matrix defined in Section 1.4.2. Using this result to (A.2.12), we obtain that

$$\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^{*}=\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right)\frac{1}{n}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(n^{-1/2}\right).$$

Since $\widehat{\boldsymbol{\theta}}^{(S)}-\boldsymbol{\theta}^{*}=\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^{*}+o_p\left(n^{-1/2}\right)$ by Lemma A.2.3, it holds that

$$\widehat{\boldsymbol{\theta}}^{(S)}-\boldsymbol{\theta}^{*}=\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right)\frac{1}{n}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(n^{-1/2}\right).$$

It can be seen that $\widehat{\boldsymbol{\theta}}^{(S)}$ and the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$ have the same influence function. Hence, under the assumptions in either part (i) or (ii) of the theorem, we have $\widehat{\boldsymbol{\theta}}^{(S)}-\widetilde{\boldsymbol{\theta}}=o_p\left(n^{-1/2}\right)$. □

Proof of Corollary 1.4.4: The result directly follows from Theorem 1.4.3 and the standard result on the asymptotic distribution of the maximum likelihood estimator. □

Proof of Theorem 1.4.5: The proof of the theorem follows from the discussion in Section 1.4.2. □

Chapter 2

EXAMINING POWER IN CLUSTERED RANDOMIZED PRICING EXPERIMENTS IN E-COMMERCE

2.1 Introduction

E-commerce platforms sell of a vast array of goods, often reaching into the thousands if not millions. To effectively manage their extensive inventory, these platforms rely on sophisticated algorithms to determine optimal pricing strategies. Given the scale and the critical role these algorithms play in the platform's operational efficiency and profitability, it is essential to rigorously evaluate their effectiveness before full-scale implementation [Hesterberg and Knight, 2024]. To this end, e-commerce platforms or other online platforms typically utilize A/B testing on the pricing policies [Cohen et al., 2016]. This approach allows them to conduct controlled experiments comparing the performance of different algorithmic versions, thereby ensuring that the most effective pricing strategy is adopted. However, subject-level randomization in these pricing experiments often encounters the challenge of treatment effect spillover due to the presence of substitute and complementary products [Bajari et al., 2023], [Berman and Van den Bulte, 2022], [Cooprider and Nassiri, 2023]. Spillover occurs when the treatment applied to one product influences the outcomes of other related products. For example a price change in one product can lead to demand shifts towards or away from its substitutes, thus affecting their sales. This spillover effect can bias the estimated treatment effect, as the control group is indirectly affected by the treatment.

To address this issue, practitioners use cluster randomized experiments (CRE), [Hayes and Moulton, 201 as a method to mitigate spillover bias. Consider all the products which bilaterally affect each others demand when either one of their prices change. If you put these products in clusters and assign the treatment randomly to these clusters , then the spillovers are prevented.

This approach helps to contain the spillover effects within clusters, reducing bias in the treatment effect estimates. However, it is important to note that empirically, the variance of treatment effect estimates in CREs is often greater than in subject-level randomization, [Bakshy et al., 2014], [Tang et al., 2010] presenting a bias-variance trade-off between subject level and cluster level randomization.

The goal of this paper is to examine power in cluster-randomized experiments by exploring how (1) different randomization schemes (matched pair cluster randomization, stratified cluster randomization, and covariate constrained cluster randomization, [Athey and Imbens, 2017]), and (2) different sets of covariate adjustments affect the power of the experiment, following the intuition layed out in [Deng et al., 2013]. Furthermore, we introduce a binned estimation as a novel approach to deal with highly skewed data, a common characteristic in e-commerce. The bins are constructed based on the pre-experiment outcome variable, and the average treatment effect is computed as the weighted average of treatment effects across bins. The weights are assigned proportionally to the inverse variance of the average treatment effect in each bin, ensuring that bins with more precise estimates are given greater emphasis in the overall calculation.

CREs introduce two additional sources of variation compared to subject-level randomization, [Crespi, 2019]: within-cluster variation and between-cluster variation. Within a cluster, observations are likely to be more similar to each other than to observations in different clusters. This correlation within clusters reduces the effective sample size and thereby the statistical power. Intuitifvely, the more the subjects within clusters are alike, the less additional information you get by having multiple subjects in the same cluster. Failure to properly account for this clustering in the analysis can lead to incorrect inferences. Specifically, if the standard errors are not adjusted for clustering, they will be underestimated, [Abadie et al., 2022], resulting in confidence intervals that are too narrow. This can inflate the Type I error rate, leading to false positive findings.

Between-cluster variation, on the other hand, captures the differences across clusters. This variation can arise from systematic differences in cluster-level characteristics, such as

product categories or price ranges. Failing to account for between-cluster variability can lead to biased estimates of the treatment effect, as the differences between clusters may confound the true impact of the intervention. One way to address between-cluster variability is by including cluster-level covariates or controls in the estimation equation, similar to the process of regression adjustment in standard regression models. By incorporating relevant cluster-level characteristics into the analysis, researchers can capture some of the systematic differences between clusters that might explain the between-cluster variability. For example, [Guo et al., 2021], uses machine learning models to do the regression adjustment, in a social network's A/B testing setting. Another work, [Taddy et al., 2015], develops a non-parametric bayesian approach to reduce the variance in online experiment setting and finds that covariate adjustment helps with precision albeit in a limited amount. This approach can improve the efficiency and precision of the estimated treatment effect by reducing unexplained variability and isolating the true treatment effect from the confounding influence of cluster-level characteristics.

However, it is important to carefully consider the choice of cluster-level covariates and ensure that they meet two key criteria. First, the covariates should have a meaningful relationship with the outcome variable and explain a significant portion of the variation in the outcome, [Rosenbaum, 1984a]. Including covariates that are not associated with the outcome will not improve the precision of the treatment effect estimate and may introduce unnecessary noise into the model. For instance, product category and price range might be relevant covariates if they are expected to have a significant impact on sales or customer behavior, while the color scheme of product images may not be as relevant. Second, the covariates should be orthogonal to the treatment assignment. It is also important to avoid adjusting for post-treatment variables that might be influenced by the treatment itself, as this can introduce bias and obscure the true treatment effect, [Rosenbaum, 1984b]. By carefully selecting cluster-level covariates that are both relevant to the outcome variable and orthogonal to the treatment, we can effectively address between-cluster variability and improve the precision and accuracy of the estimated treatment effect in clustered randomized

trials.

The rest of the paper proceeds as follows. Section 2.2 presents a model that shows how CREs increase variance in treatment effect estimates. Section 2.3 describes the randomization schemes, balance checks, and introduces a binned estimator for high-kurtosis data. Section 2.4 explains the simulation setup, data generation process, and presents results from different randomization methods. Section 2.5 concludes with implications for practice.

2.2 Model

In this part, we demonstrate the extra variance caused by CRE, in line with the work of [Crespi, 2019].

We presume that each cluster possesses its own mean, and the units within the cluster have outcomes that fluctuate around that mean. The model for the outcome of individual i in cluster j, represented as Y_{ij}, is

$$Y_{ij} = \mu_j + \epsilon_{ij} \tag{2.2.1}$$

where μ_j denotes the mean for cluster j and ϵ_{ij} represents the error term, which shows the difference between the unit's observed outcome Y_{ij} and the cluster mean outcome μ_j.

We also assume that our clusters are taken from a population of clusters that has an overall mean, with the cluster means varying around it. The model for the mean of cluster j, μ_j, is

$$\mu_j = \alpha + u_j \tag{2.2.2}$$

where α stands for the population mean, which is assumed to be fixed, and u_j is a random effect that represents the difference between cluster j's mean and the population mean. Combining these we have,

$$Y_{ij} = \alpha + u_j + \epsilon_{ij}. \tag{2.2.3}$$

The error terms are presumed to be normal with $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and they are assumed to be independent of each other.

Then, the total variance of an observation Y_{ij}, not conditional on cluster, can be decomposed as the sum of two independent variance components, one at the cluster level and the other at the individual level,

$$Var(Y_{ij}) = \sigma_y^2 = \sigma_u^2 + \sigma_\epsilon^2 \tag{2.2.4}$$

The proportion of the total variance that is due to clustering is,

$$\rho = \frac{\sigma_u^2}{\sigma_y^2} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} \tag{2.2.5}$$

It is shown(also in appendix B.1) that ρ also equals to the correlation between two different observations from the same cluster and hence called Intra-cluster Correlation Coefficient(ICC).

$$ICC = Corr(Y_{ij}, Y_{kj}) = \rho \tag{2.2.6}$$

Suppose each of n_2 clusters we have, has n_1 observations each. Then the sample mean for cluster j is,

$$\bar{Y_j} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{ij}$$

And its variance is shown to be:

$$Var(\bar{Y_j}) = \frac{1}{n_1}\Big[Var(Y_{ij}) + (n_1 - 1)Cov(Y_{ij}, Y_{kj})\Big]$$

$$= \frac{\sigma^2}{n_1}\Big[1 + (n_1 - 1)\rho\Big]$$

Similarly, let $\bar{Y}$ denote overall sample mean over all observations , $\bar{Y} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} Y_{ij}$.

Since observations in different clusters are assumed to be independent the variance of $\bar{Y}$ is,

$$Var(\bar{Y}) = \frac{\sigma_y^2}{n_1 n_2}\Big[1 + (n_1 - 1)\rho\Big]$$

If the total number of observations, $n_1 n_2$ had been independent the second term would have been 0.

Suppose now that our n_2 clusters are randomized to treatment and control, with $\frac{n_2}{2}$ clusters in each. To accommodate different population means in the treatment and control, we modify the model for the mean of cluster j to be:

$$\mu_j = \alpha + \tau w_j + u_j$$

Thus, α is the population mean, and τ is the difference in means between the treatment and control. Note that the treatment indicator w_j is subscripted only by j and not by i, since treatment is assigned at the cluster level. The single equation model for the outcome Y_{ij} is:

$$Y_{ij} = \alpha + \tau w_j + u_j + \epsilon_{ij} \tag{2.2.7}$$

The total variance of an observation is still $Var(Y_{ij}) = \sigma_u^2 + \sigma_\epsilon^2$ and we still express the ICC as $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$, although it is possible that the magnitude of the variance components differs between the treatment and control.

Our interest is to estimate the treatment effect τ. An unbiased estimate of τ can be obtained as the difference of means between treatment and control. Indexing the treatment by $W = 0, 1$ and the outcomes as Y_{ij}^W, our estimate of the treatment effect is $\hat{\tau} = \bar{Y}^1 - \bar{Y}^0$ with variance:

$$Var(\hat{\tau}) = \frac{4\sigma_y^2}{n_1 n_2}\left[1 + (n_1 - 1)\rho\right]$$

The test for a treatment effect is the test of the null hypothesis $H_0 : \tau = 0$ in the

treatment and control. This test can be conducted using the test statistic:

$$T = \frac{\hat{\tau}}{SE(\hat{\tau})}$$

where $SE(\hat{\tau}) = \sqrt{Var(\hat{\tau})}$.

2.3 Methodology

In this section, we look at a comprehensive survey of techniques widely used in the technology sector, including various randomization schemes, balance checks, estimating equation, as well as our novel binned estimator. We focus on the statistical power of the experiment to quantify the precision loss of clustered randomized experiments. Power is calculated using the following formula, [Cohen, 1988]:

$$\text{Power} = 1 - \Phi\left(z - \frac{\theta_{pct} \cdot \hat{\mu}_{y_0} \cdot \sqrt{2SR}}{\hat{\sigma}_{\hat{\tau}}}\right) \tag{2.3.1}$$

where θ_{pct} denotes the percentage of the effect, $\hat{\mu}_{y_0}$ is the mean of the outcome variable in the control group, Φ is the cumulative distribution function of the standard normal distribution, SR is the sampling ratio, z is the $(1 - \alpha/2)$ standard normal quantile, and $\hat{\hat{\tau}}$ is the estimated standard error of the treatment effect.

In matched pair cluster randomization, clusters are matched into pairs based on similar characteristics or covariates that are thought to be related to the outcome of interest. Once the pairs are formed, one cluster within each pair is randomly assigned to the treatment group, while the other is assigned to the control group. This scheme helps to balance the treatment and control groups in terms of the matched covariates, which can increase the power of the study.

In stratified cluster randomization, clusters are first divided into strata based on one or more covariates that are believed to be related to the outcome. Within each stratum, clusters are then randomly assigned to the treatment or control group. This scheme ensures that the treatment and control groups are balanced within each stratum, which can increase the

power of the study.

In covariate constrained cluster randomization, the randomization of clusters to treatment and control groups is performed in a way that minimizes the imbalance in the distribution of one or more covariates between the groups. This is typically done using optimization algorithms that search for a randomization scheme that satisfies the balance constraints.

We employ two balance checks that are performed to ensure that the treatment and control groups are indeed balanced on the relevant covariates. We use a t-test for mean equivalence and the Kolmogorov-Smirnov (KS) test [Massey, 1951] for distributional equivalence. The t-test compares the means of the covariates between the treatment and control groups, testing the null hypothesis that the means are equal. If the p-value from the t-test is greater than a pre-specified significance level (e.g., 0.05), we fail to reject the null hypothesis and conclude that the means are not significantly different, indicating balance between the groups. The KS test, on the other hand, compares the entire distribution of the covariates between the treatment and control groups. It tests the null hypothesis that the two distributions are the same. Similarly, if the p-value from the KS test is greater than the significance level, we fail to reject the null hypothesis and conclude that the distributions are not significantly different, indicating balance.

For a given clustering method, randomization scheme, and set of controlling variables we estimate the treatment effect, its standard error and quantify the variance in the estimate by focusing on power of the test. In practice, if researchers has data of previous experiments, they can estimate the standard error of the treatment effect $\hat{SE}(\hat{\tau})$ by conducting A/A tests using pre-experiment data. By construction, the true treatment effect is zero. The treatment effect and its standard error are estimated using the following linear regression adjustment specification:

$$y_{ig,post} = \alpha + \tau W_g + X_{ig,pre} + \epsilon_{ig} \tag{2.3.2}$$

where $y_{ig,post}$ is the outcome variable for individual i in group g post-intervention, W_g

is the treatment indicator for group g, $X_{ig,pre}$ are pre-intervention covariates, and ϵ_{ig} is the error term. Standard errors are estimated using traditional cluster-robust standard error estimates [Arellano, 1987].

2.3.1 High Kurtosis and a Binned Estimator

This section explores the issue of high-kurtosis data, which is common in e-commerce settings and why it is problematic in experimental settings, examining its impact on key statistical measures, randomization schemes, and analytical techniques.

Kurtosis, defined as $\kappa = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$, where μ is the mean and σ^2 is the variance, quantifies the "tailedness" of a probability distribution. High kurtosis indicates a higher probability of extreme values or outliers, which fundamentally affects the reliability of key statistical estimators. The sample mean $\hat{\mu}$ becomes less stable, as its variance $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$ is inflated due to the increased σ^2 resulting from heavy tails. Similarly, the sample variance $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$ becomes less reliable, with its variance increasing as $\text{Var}(\hat{\sigma}^2) \approx \frac{\sigma^4(\kappa-1)}{n}$ for large n. This instability prevents researches to have robust experiment results since even in the simplest form of a treatment effect estimate, the variance of the treatment effect is: $\text{Var}(\tau) = \text{Var}(\hat{\mu}_T) + \text{Var}(\hat{\mu}_C) = \frac{\sigma_T^2}{n} + \frac{\sigma_C^2}{n}$.

In the context of balancing randomization schemes, high kurtosis undermines the effectiveness of common approaches. Matched pair randomization wouldn't work as intended since the extreme values dominate the matching process, leading to unstable distances between units and unreliable within-pair estimates. Stratified randomization suffers from skewed stratification and increased within-stratum variance, reducing the benefits of this approach. Covariate constrained randomization is similarly affected, with unstable means leading to ineffective balance metrics. Failure to address these challenges can lead to incorrect conclusions and compromised research integrity.

Binned Estimator

We introduce a binned estimator designed to reduce the variance of treatment effect estimates in the presence of high-kurtosis data, while maintaining the cluster structure. The approach first partitions the data into bins based on pre-experiment outcomes, then applies cluster randomization within each bin.

Let Y be the pre-experiment outcome variable. We first partition the product space into B bins based on the quantiles of Y. For each bin $b = 1, ..., B$, we observe clusters $g = 1, ..., G_b$, where G_b denotes the number of clusters in bin b. Let n_{gb} denote the number of products in cluster g of bin b.

Within each bin b, we estimate the treatment effect using:

$$Y_{igb} = \alpha_b + \tau_b W_{gb} + \epsilon_{igb}$$

where Y_{igb} is the outcome for product i in cluster g of bin b, W_{gb} is the treatment indicator for cluster g in bin b, τ_b is the bin-specific treatment effect and ϵ_{igb} represents the error term.

The bin-specific treatment effect estimator $\hat{\tau}_b$ is computed with cluster-robust standard errors to account for within-cluster correlation. The overall average treatment effect estimator is then:

$$\hat{\tau} = \sum_{b=1}^{B} w_b \hat{\tau}_b$$

where the weights are inversely proportional to the cluster-robust variance estimates:

$$w_b = \frac{1/\hat{v}_b}{\sum_{k=1}^{B} 1/\hat{v}_k}$$

The cluster-robust variance estimator $\hat{v}_b$ for each bin follows [Arellano, 1987], computed as:

$$\hat{v}_b = (X_b' X_b)^{-1} \left(\sum_{g=1}^{G_b} X_{gb}' \hat{\epsilon}_{gb} \hat{\epsilon}_{gb}' X_{gb} \right) (X_b' X_b)^{-1}$$

where X_b is the design matrix for bin b containing a constant and treatment indicator, X_{gb}

contains the regressors for cluster g in bin b, and $\hat{\epsilon}_{gb}$ represents the vector of estimated residuals for cluster g in bin b.

Several assumptions underpin this estimation approach. First, we assume that cluster boundaries largely align with bin boundaries, meaning most clusters have their members concentrated within a single bin. While this assumption may appear restrictive, it often holds in practice as products within the same cluster (e.g., similar products or substitutes) tend to have similar pre-experiment outcomes. Second, we assume that within each bin, the cluster structure remains stable over the experimental period, allowing for consistent estimation of cluster-robust standard errors.

The effectiveness of this binned estimator depends crucially on the bin construction. We recommend forming bins based on pre-experiment outcome quantiles, with the number of bins chosen to balance the trade-off between variance reduction and estimation precision. Too few bins may fail to adequately address the high-kurtosis problem, while too many bins could result in imprecise bin-specific treatment effect estimates due to smaller sample sizes.

One limitation of this approach is when clusters significantly span multiple bins. In such cases, the independence assumption between bin-specific treatment effects may be violated, potentially leading to underestimated standard errors in the overall treatment effect. Researchers should carefully examine their data structure and potentially consider alternative binning strategies when cluster-bin overlap is substantial. Additionally, while this method effectively addresses the high-kurtosis problem, it may introduce some bias in the overall treatment effect estimate if treatment effects are heterogeneous across bins and the weighting scheme differs substantially from the population distribution.

Despite these limitations, the binned estimator provides a practical solution to the challenges posed by high-kurtosis data in cluster randomized experiments. By combining the variance-reduction benefits of binning with proper accounting for cluster-level randomization, this approach allows for more reliable inference in e-commerce experimentation settings where extreme values are common.

2.4 Simulation Analysis

2.4.1 Data Generating Process

This process involves generating price and quality tensors, applying a treatment to prices post-randomization, simulating customer demand based on price and quality changes, and finally calculating revenues and profits for each product. The design of the simulation allows for careful control of variables and the systematic exploration of treatment effects, all while incorporating noise and random fluctuations to approximate real-world conditions. Throughout, we use t to index date, j to index cluster, i to index products. We use variable hyperparameters (n_1, n_2) to denote product per cluster and number of clusters respectively.

DGP: Price and Quality Tensors

The first step in the simulation is to generate a quality matrix for each product across clusters. Specifically, the quality matrix, Q_{n_1,n_2}, is drawn from a uniform distribution $U(\mu_j - 15, \mu_j + 15)$, where the mean quality μ_j for each cluster is sampled from another uniform distribution $U(5, 99)$. This creates product qualities that vary across clusters but stay within a predefined range, simulating realistic variability in product quality across different clusters.

Following the generation of quality values, price fluctuations are introduced. These fluctuations are modeled as a normal distribution $N(0, 1)$ for each product over time, capturing the natural variability in prices that might arise due to market conditions, competition, or other external factors. Specifically, for each date-product-cluster combination, the price error term $PriceErr_{t,i,j}$ is drawn from this normal distribution, adding random noise to the base quality-derived prices.

The final price tensor, $P_{t,i,j}$, is then constructed by adding the price error to the quality values, i.e., $P_{t,i,j} = Q_{i,j} + PriceErr_{t,i,j}$. This approach integrates product quality and price fluctuations into a cohesive structure, allowing for rich heterogeneity across products and clusters.

Once the baseline prices have been set, the treatment is applied. The treatment effect

consists of increasing prices by a constant percentage, for example 10%, after a predefined trigger date, such that for any $t > triggerDate$, the treatment-adjusted price becomes $p_{t,i,j} \times 1.1$. This manipulation simulates the effect of a price change intervention (such as a price hike or discount) that might be tested in an experiment. By structuring the treatment in this way, the simulation ensures that any observed changes in demand, revenue, or profits can be causally linked to the price adjustment.

DGP: Demand Simulation

Once the price and quality tensors have been generated and the treatment applied, the next step is to simulate customer demand. This is achieved through the use of a utility function that models customer preferences. The utility for product n_1 in cluster n_2 on day nd is given by:

$$U_{t,i,j} = \alpha \cdot P_{t,i,j} + \beta \cdot Q_{i,j} + e_{t,i,j}$$

Here, α and β are coefficients representing the sensitivity of customers to price and quality, respectively, while $e_{t,i,j}$ is Gumbel distributed error terms captures any idiosyncratic factors affecting utility that are not directly modeled. This formulation assumes that customer utility is influenced by both the price and quality of the product, with random noise reflecting unobserved variations in preferences.

Since the error term is Gumbel, the optimal choice probabilities are represented by a softmax function of the logits for each day-product-cluster combination, as per [McFadden, 1974]:

$$p_{t,i,j} = \text{softmax}(\alpha \cdot P_{t,i,j} + \beta \cdot Q_{i,j})$$

This probabilistic choice model allows for a more flexible and realistic simulation of customer behavior, where higher utility leads to a higher likelihood of choosing a particular product, but other options are still possible due to the random component $e_{t,i,j}$.

Customer choices are then simulated by drawing samples based on the calculated prob-

abilities. Each day, 500 customers are simulated, with each customer choosing a product based on the softmax probabilities.

DGP: Revenues and Profits

Finally, the revenues and profits for each product in each cluster are calculated based on the simulated prices and customer choices. The cost for producing each product, C_{nd,n_1,n_2}, is assumed to be proportional to the product's quality, with a cost multiplier applied to the quality matrix, i.e., $C_{t,i,j} = \text{Cost Markup} \cdot Q_{n_1,n_2}$. This assumption captures the intuition that higher-quality products are more expensive to produce.

Revenues are calculated as the product of the price and the number of customers who chose the product, i.e.,

$$R_{t,i,j} = P_{t,i,j} \cdot \text{Counts}_{t,i,j}$$

where $\text{Counts}_{t,i,j}$ represents the number of customers who chose product i in cluster j on day t. This measure of revenue provides insight into how changes in price and demand influence the overall financial performance of each product.

Profits, on the other hand, are calculated by subtracting the production cost from the revenue for each product:

$$\Pi_{t,i,j} = (P_{t,i,j} - C_{t,i,j}) \cdot \text{Counts}_{t,i,j}$$

Data

To have a feel of what the generated data looks like, let's examine some key visualizations that illustrate the characteristics of our simulated e-commerce environment. Figure B.1 presents two heatmaps illustrating the average quality and average price distributions across a sample of five clusters. The heatmaps provide a visual representation of the relationship between product quality and pricing within the simulated environment. In the quality heatmap,

clusters 3 and 4 exhibit notably higher average quality levels compared to the other clusters, as indicated by the darker shading. This variation in quality across clusters aligns with the data generation process, where cluster-specific mean qualities are drawn from a uniform distribution. The price heatmap closely mirrors the quality distribution, reflecting the price-setting mechanism in the simulation where prices are derived from quality values with added random fluctuations. The similarity between the quality and price heatmaps demonstrates the strong correlation between these two variables in the simulated market environment. This visualization effectively captures the heterogeneity in product characteristics across clusters.

Figures B.2a, B.2b and B.2c depict the count paths for products in Cluster 1 under three distinct treatment conditions: no treatment, subject-level randomization, and cluster-level randomization, respectively. These figures provide insight into the impact of different treatment applications on product demand over time. Figure B.2a shows the baseline scenario with no treatment applied. The count paths exhibit natural fluctuations over time, reflecting the stochastic nature of the demand simulation process. In Figure B.2b, where subject-level randomization is applied, the divergence between treated (red lines) and untreated (blue lines) products becomes apparent after the treatment implementation at day 30. This divergence illustrates the direct effect of the price increase on individual product demand, as well as potential spillover effects within the cluster. Figure B.2c presents the scenario under cluster-level randomization, where all products within the cluster receive the treatment (indicated by red lines). The uniform application of the treatment across all products in the cluster allows for a clear observation of the aggregate impact on demand within the cluster with some of the demand shifts to the outside good. These figures demonstrate the different impacts of treatment application methods on product demand, highlighting the interference effects in subject-level randomization and the cluster-wide impact in cluster-level randomization.

Figures B.3a, B.3b and B.3c illustrate the price paths for products in Cluster 1 under the same three treatment conditions as in Figure 2: no treatment, subject-level randomization, and cluster-level randomization, respectively. These figures provide a visual representa-

tion of how the different treatment applications affect product pricing over time. Figure B.3a displays the baseline scenario with no treatment applied. The price paths show minor fluctuations over time, reflecting the random price errors incorporated in the data generating process. In Figure B.3b, which depicts subject-level randomization, a clear distinction emerges between treated (red lines) and untreated (blue lines) products after day 30, the point of treatment implementation. The treated products exhibit a noticeable upward shift in price, consistent with the 10% price increase specified in the treatment. Figure B.3c presents the cluster-level randomization scenario, where all products within the cluster receive the treatment (indicated by red lines). The uniform upward shift in prices across all products after day 30 clearly illustrates the cluster-wide application of the price increase treatment. These figures effectively capture the direct impact of the treatment on product pricing under different randomization schemes. They provide a clear visual representation of how the data generating process we descibed previously, and implements the price increase treatment, which is crucial for understanding the subsequent effects on demand and revenue in the simulated e-commerce environment.

Discussion

The use of simulated data in our study of cluster randomized experiments (CREs) offers both advantages and limitations. One significant benefit is the tractability it provides in examining DGP, interference effects, standard errors of treatment effect, the power of the experiment and lack thereof. In real-world data, isolating and quantifying these can be challenging due to the complex interplay of numerous variables and lack of ground truth. However, our simulated environment allows us to precisely control and observe how changes in one product's price affect the demand for related products.

Another advantage of using simulated data is the flexibility it offers in testing various scenarios, including edge cases that might be rare in real-world data. For instance, we can easily generate datasets with high kurtosis, a characteristic often observed in e-commerce data. This flexibility enables us to rigorously test our binned estimator under a wide range

of conditions, ensuring its robustness.

However, the use of simulated data obviosly not optimal because it is not real world data and it presents certain limitations. One potential drawback is that simulated environments may not fully capture the complexity and nuances of real-world e-commerce ecosystems. While our simulation incorporates key factors such as price elasticity and product quality, it may still miss subtle interactions or external influences that could impact actual consumer behavior. This simplification, while necessary for tractability, might lead to overly optimistic assessments of our methods' effectiveness.

Furthermore, in real-world scenarios there are huge number of covariates which is hard to replicate. In actual e-commerce platforms, a vast number of variables - from product characteristics, product interactions, and seasonal trends - can influence purchasing decisions. Our simulated environment, while sophisticated, cannot fully replicate this multidimensional complexity.

Despite these limitations, the insights gained from our simulated study provide a robust foundation for understanding the behavior of CREs in e-commerce settings and the potential improvements offered by our binned estimator.

2.4.2 Balance Checks

Balance in cluster randomized experiments is crucial for maximizing the power of the experiment. Balanced treatment and control groups ensure that any observed differences in outcomes can be attributed to the treatment effect rather than pre-existing differences between groups. This balance reduces the variance in treatment effect estimates, leading to more precise measurements and increased statistical power. We employed two statistical tests to assess balance between treatment and control groups, as outlined in the section 2.3. The t-test for mean equivalence examines the null hypothesis that the means of the variables are equal between the treatment and control groups, against the alternative that they differ. The Kolmogorov-Smirnov (KS) test for distributional equivalence, as described by [Massey, 1951], tests the null hypothesis that the distributions of the variables are identical

between the two groups, with the alternative being that they differ in some way.

Examining the results across different randomization methods, we observe varying degrees of balance achieved. The simple subject-level randomization shows remarkably good balance across all variables, with high p-values (all > 0.4) for both t-tests and KS tests, indicating no significant differences in means or distributions between treatment and control groups. This aligns with our expectations, as randomization at the product level tends to distribute products from all clusters evenly between groups.

In contrast, the simple cluster-level randomization exhibits poorer balance, where we see statistically significant differences ($p < 0.05$) in both means and distributions. This highlights a key challenge in CREs, where randomizing entire clusters can lead to imbalances due to between-cluster variations, as discussed in our model section.

The matched-pair cluster randomization shows a marked improvement over simple cluster randomization, with high p-values (all > 0.4) across all variables, indicating good balance. This method's effectiveness stems from its ability to pair similar clusters before randomization, mitigating the impact of between-cluster differences.

Stratified cluster randomization also demonstrates good balance, though not as consistently as the matched-pair method, with p-values ranging from 0.26 to 0.85. The covariate constrained cluster randomization shows similar results to the stratified method, with no significant differences detected, but with slightly lower p-values overall compared to the matched-pair approach.

While subject-level randomization achieves the best balance, it may not always be feasible due to spillover effects, as discussed in our introduction. Among the cluster-level methods, matched-pair randomization appears to be the most effective in achieving balance, followed closely by stratified and covariate constrained methods.

2.4.3 Results

The results of the numerical simulation exercise are presented in Tables I for moderate Kurtosis. First, we can observe the decrease in statistical power when transitioning from

subject-level to cluster-level randomization. This aligns with our earlier discussion on the additional sources of variation introduced by CREs. For instance, in Table I, we see the power drop from 0.85 for simple subject-level randomization to 0.32 for simple cluster-level randomization without control variables. This decrease in power underscores the challenges inherent in CREs and the importance of employing strategies to mitigate this loss of precision.

One such strategy, as predicted by the CUPED (Controlled-experiment Using Pre-experiment Data), [Deng et al., 2013] method, is the inclusion of pre-experiment outcome variables as controls. Our results strongly support the effectiveness of this approach. Across all randomization methods, controlling for pre-experiment profit consistently leads to substantial increases in power. For example, in the case of simple cluster-level randomization in Table I, power increases from 0.32 to 0.59 when pre-experiment profit is included as a control variable.

Furthermore, our results demonstrate that including cluster-level covariates in addition to pre-experiment outcomes can further enhance power. We observe an additional increase in power of approximately 5-10 percentage points when cluster-level variables are included. This finding supports our earlier discussion on the importance of addressing between-cluster variability through the inclusion of relevant cluster-level characteristics in the estimation equation.

Interestingly, while different randomization methods show some variation in performance, their impact on power is relatively modest compared to the effects of including control variables. Among the cluster-level randomization methods, matched-pair and stratified randomization tend to perform slightly better than simple cluster-level randomization, particularly when combined with appropriate control variables. This aligns with our earlier discussion on the balance checks, where these methods demonstrated improved balance compared to simple cluster-level randomization.

It's also worth noting the impact of data characteristics on these results. High kurtosis generally leads to lower power across all methods, as described in [Wilcox, 2012]. This observation underscores the challenges posed by high-kurtosis data in e-commerce settings,

as discussed in our methodology section, and highlights the importance of developing robust estimation techniques like our proposed binned estimator to address these issues. The binned estimator's power is very high across all randomizations. This is because of the inverse variance weighting which effectively ignores the bins that have very high intrinsic variance due to the tailedness of the data. Admittedly this introduces some bias to estimator especially if the treatment effects accross bins are different. Practically, the most robust approach is to make an A/B test for each product category or bin separately and get the interpret the incrementality of the algorithm maintaning reasonable kurtosis in the data. Binned estimator is an approach where this is not feasible or practical.

Table I: Outcome variable: Profit, with a kurtosis of 10.40, comparing various randomization methods and their impact on standard error, z-statistic, p-value, and statistical power.

Randomization	**Control Vars**	**SE**	**z**	**p**	**Power**
Simple Subject Level Rando.	-	39.27	-21.36	0	0.85
Simple Subject Level Rando.	pre_profit	48.44	-17.31	0	0.68
Simple Subject Level Rando. Binned	-	18.71	-19.94	0	0.99
Simple Cluster Level Rando.	-	54.79	-2.30	0.003	0.32
Simple Cluster Level Rando.	pre_profit	36.94	-6.45	0	0.59
Simple Cluster Level Rando.	pre_profit, pre_cluster_vars	35.85	-6.52	0	0.62
Simple Cluster Level Rando. Binned	-	24.10	-4.44	0	0.99
Matched-Pair Cluster Level Rando.	-	55.25	-4.18	0	0.37
Matched-Pair Cluster Level Rando.	pre_profit	37.10	-6.12	0	0.69
Matched-Pair Cluster Level Rando.	pre_profit, pre_cluster_vars	36.12	-6.36	0	0.71
Matched-Pair Cluster Level Rando. Binned	-	21.22	-5.36	0	0.99
Stratified Cluster Level Rando.	-	54.94	-4.12	0	0.36
Stratified Cluster Level Rando.	pre_profit	34.41	-7.27	0	0.72
Stratified Cluster Level Rando.	pre_profit, pre_cluster_vars	35.60	-6.98	0	0.69
Stratified Cluster Level Rando. Binned	-	23.09	-5.39	0	0.99
Covariate Constrained CL Rando.	-	56.79	-4.05	0	0.35
Covariate Constrained CL Rando.	pre_profit	36.05	-6.45	0	0.69
Covariate Constrained CL Rando.	pre_profit, pre_cluster_vars	34.91	-6.52	0	0.73
Covariate Constrained CL Rando. Binned	-	25.07	-3.75	0	0.99

2.5 Conclusion

In this work we addressed a pervasive challenge in e-commerce experimentation: the heightened variance in cluster randomized experiments (CREs). While CREs are essential for mitigating spillover effects in pricing experiments, they introduce additional variance in treatment effect estimates, potentially compromising statistical power. We developed a comprehensive methodological framework to analyze this trade-off and explore various approaches to improve the precision of estimates.

Our analysis, based on simulated e-commerce data, demonstrated that the transition from subject-level to cluster-level randomization substantially reduces statistical power, with power dropping from 0.85 to 0.32 in our baseline scenario. However, we found that this loss in precision can be partially mitigated through several strategies. The inclusion of pre-experiment outcomes as control variables consistently improved power across all randomization methods, with further gains achieved by incorporating cluster-level covariates. Among the cluster-level approaches, matched-pair and stratified randomization showed slightly better performance than simple cluster-level randomization, particularly when combined with appropriate controls.

The simulation results also highlighted the particular challenges posed by high-kurtosis data, which is common in e-commerce settings. While we introduced a binned estimator as a potential approach to handle such data, further research is needed to fully understand its properties and limitations. Our findings suggest that the most practical approach may be to conduct separate analyses for different product categories or groups of products when feasible. We underscore the importance of careful experimental design and the strategic use of control variables to maximize the precision of treatment effect estimates in cluster randomized experiments.

Chapter 3

TWO STAGE TEXT REGRESSION USING TRANSFORMER-BASED ENCODINGS

3.1 Introduction

The field of economics has witnessed a paradigm shift in recent years, marked by an increasing ability to harness diverse and unconventional data sources. This evolution has been particularly pronounced in the domain of text analysis, where the integration of textual data into econometric models has gained significant traction. The proliferation of digital text across various platforms has opened up new avenues for economic research, offering rich insights into human behavior, market dynamics, and societal trends.

This paper introduces a novel two-stage text regression methodology that leverages transformer-based encodings to enhance the integration of textual data in econometric models, demonstrating its effectiveness in capturing gender-associated language patterns on an online economics forum. Our approach aims to bridge the gap between the vast potential of textual information and the rigorous quantitative frameworks that underpin economic analysis.

The use of text data in economics, while still in its nascent stages, has already demonstrated considerable promise across various subfields. In finance, for instance, [Baker and Wurgler, 2006b] pioneered the application of sentiment analysis to explain cross-sectional variations in stock returns. More recently, [Chen et al., 2021] explored how hedge funds can leverage sentiment forecasts to generate superior returns. Beyond finance, [Taddy, 2015b] applied text regression to Yelp reviews, while [Gentzkow et al., 2019b] developed a framework to measure group differences in high-dimensional choices, applying it to gauge polarization in U.S. Congressional speeches.

Despite these advancements, the most common approach to incorporating text data in econometric models remains to be the bag-of-words method. This approach, while straightforward and computationally efficient, treats text as an unordered collection of words, disregarding syntax and context. Its primary advantages lie in its simplicity and interpretability, as it allows for direct analysis of word frequencies and their relationships with outcome variables. However, the bag-of-words approach suffers from significant limitations, including its inability to capture word order, context, or semantic relationships, potentially leading to loss of crucial information embedded in the text.

Our proposed methodology builds upon and significantly extends the work of [Wu, 2018], who employed text analysis and machine learning techniques to examine gendered language on an online economics forum. While the approach of [Wu, 2018] utilized a Lasso-regularized logistic regression model based on word frequencies, our method leverages advanced natural language processing techniques, specifically transformer-based encodings introduced in the seminal paper [Vaswani et al., 2017] to capture richer semantic information.

The key innovation of our approach lies in its two-stage framework. In the first stage, we employ dimensionality reduction techniques based on deep learning models (e.g., [Devlin et al., 2018, Brown et al., 2020]) to represent textual data in a lower-dimensional space. This allows us to capture complex linguistic patterns, including word order, long-range dependencies, and contextual nuances, which are often lost in simpler representations. The second stage involves estimating a model to infer the association between the outcome variable and the information contained within the text data, as represented by these lower-dimensional encodings.

This two-stage approach offers several advantages over existing methods. Firstly, it enables a more nuanced and context-aware analysis of text data, potentially uncovering subtle patterns that might be missed by bag-of-words or similar approaches. Secondly, the reduction in dimensionality from the full vocabulary size to a more manageable embedding size enhances model tractability and reduces the risk of overfitting. Finally, the modular nature of our approach allows for easy integration of state-of-the-art text encoders and flexibility in the choice of the second-stage econometric model.

While bag-of-words remains the most common approach in economic applications, recent research has begun exploring more sophisticated natural language processing techniques. Notably, [Vafa et al., 2023] introduced CAREER, a transformer-based model for predicting job sequences that employs a two-stage approach: first pre-training on large-scale resume data to learn general career patterns, then fine-tuning on smaller, carefully constructed longitudinal survey datasets. Their success in combining transformer architectures with economic data for sequence prediction tasks demonstrates the potential of modern NLP methods in economics. Building on this direction, our methodology extends the application of transformers to a broader class of economic problems through a novel two-stage text regression framework that can capture rich semantic information in various economic contexts.

The flexibility of our approach opens up a wide array of potential applications in economic research. For instance, in monetary policy analysis, our method could be applied to predict interest rate decisions based on the text reports published by Federal Reserve after the Federal Open Market Committee (FOMC) meetings. By encoding the nuanced language used by committee members, the model could potentially capture subtle shifts in sentiment or policy leanings that might not be apparent in more traditional quantitative indicators.

In consumer behavior literature, our methodology could significantly enhance the analysis of online reviews. Unlike simple sentiment analysis, our approach could potentially uncover complex relationships between specific aspects of product descriptions and consumer satisfaction, accounting for context and idiomatic expressions that might be lost in simpler text representations.

Furthermore, in labor economics, our method could be applied to large-scale survey data to identify emerging trends in job satisfaction, work-life balance, or skill demands. By capturing the richness of open-ended responses, our approach could provide more nuanced insights than those derived from traditional categorical survey questions. In this paper we aim to showcase the potential of our approach in advancing the integration of text analysis into economic research, opening new pathways for understanding the complex interplay between language, behavior, and economic outcomes.

This paper contributes to the growing literature on text-as-data in economics by introducing a novel two-stage text regression framework that seamlessly integrates state-of-the-art transformer embeddings into standard econometric models. First, by moving beyond traditional bag-of-words methods, it brings deep contextual understanding of language into economic analysis, enabling the capture of subtle linguistic signals that would otherwise remain undetected. Second, by removing explicit identifiers and focusing on contextual clues, the approach clarifies how even subtle textual features can encode group distinctions—offering improved robustness over dictionary-based classification. Finally, this methodology is general enough to be adapted for various economic domains, from policy evaluation to behavioral experiments, thus broadening the applicability of modern NLP tools for economists. By systematically demonstrating how rich text embeddings can be combined with econometric techniques, the paper lays a groundwork for future research seeking either causal or predictive insights from unstructured text.

In Section 3.2, we provide an overview of the transformer-based architectures used to encode text into numerical representations, focusing on how these embeddings capture rich semantic information. Section 3.3 outlines our two-stage regression framework, which utilizes these embeddings as inputs to a variety of second stage models to estimate the relationship between text and the target variable. Sections 3.4 and 3.5 presents an empirical application of this methodology to data from [Wu, 2018]. Finally, Section 3.6 concludes.

3.2 Mathematical Interpretation of Transformers

The theoretical foundation of our embedding approach rests on recent advances in the mathematical analysis of transformer architectures. [Yun et al., 2020] proved that transformers are universal approximators of continuous sequence-to-sequence functions, providing theoretical guarantees for their representational power. This result ensures that our first stage embeddings can capture the complex semantic relationships present in the text data.

Let document i be represented as a sequence of T tokens: $s_i = [w_1^i, w_2^i, \ldots, w_T^i]$, where w_t^i represents the t-th token in document i. Each token w_t^i is first mapped to a vector in $\mathbb{R}^d$

through an embedding matrix $E \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the vocabulary size. This gives us our initial representation:

$$Z^{(0,i)} = [E(w_1^i); E(w_2^i); \ldots; E(w_T^i)] \in \mathbb{R}^{T \times d}$$

The transformer processes these representations through L layers. For each layer $l \in \{1, \ldots, L\}$, $Z^{(l,i)} \in \mathbb{R}^{T \times d}$ represents the sequence of token embeddings where each row $z_t^{(l,i)} \in \mathbb{R}^d$ is the embedding of token t. The transformation from $Z^{(l-1,i)}$ to $Z^{(l,i)}$ occurs through two sequential sub-layers: multi-head self-attention mechanism and position-wise feed-forward networks.

Specifically, for each sequence i, at each layer l, the token representations are projected into query, key, and value spaces:

$$Q^{(l,i)} = Z^{(l-1,i)} W_Q, \quad K^{(l,i)} = Z^{(l-1,i)} W_K, \quad V^{(l,i)} = Z^{(l-1,i)} W_V$$

where $Q^{(l,i)}, K^{(l,i)}, V^{(l,i)} \in \mathbb{R}^{T \times d_k}$ and d_k represents the dimension of the query and key vectors in the attention mechanism.[1] $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable parameters.

For each attention head, after the projections, the attention weights for sequence i at layer l are computed as:

$$A^{(l,i)} = \text{softmax}\left(\frac{Q^{(l,i)}(K^{(l,i)})^\top}{\sqrt{d_k}}\right)$$

$A^{(l,i)} \in \mathbb{R}^{T \times T}$ is the attention matrix where each entry $\alpha_{ts}^{(l,i)}$ represents the attention weight from token t to token s.[2]

For each head, the representations are updated as:

$$H^{(l,i)} = A^{(l,i)} V^{(l,i)}$$

---

[1] Typically $d_k = d/h$ for h attention heads.

[2] The softmax is applied row-wise, ensuring $\sum_{s=1}^{T} \alpha_{ts}^{(l,i)} = 1$ for all t

where $H^{(l,i)} \in \mathbb{R}^{T \times d_k}$ contains the contextually updated representations.

Note that the process described above is repeated h times with different learned projections, yielding:

$$\text{MultiHead}(Z^{(l-1,i)}) = \text{Concat}(H_1^{(l,i)}, ..., H_h^{(l,i)})W_O$$

where $W_O \in \mathbb{R}^{(h \cdot d_k) \times d}$.

These multi-head attention operations are followed by position-wise feed-forward networks, yielding the final update for layer l:

$$Z^{(l,i)} = \text{FFN}(\text{MultiHead}(Z^{(l-1,i)}))$$

where FFN applies the same feed-forward transformation to each position independently:

$$\text{FFN}(Y) = \max(0, YW_1 + b_1)W_2 + b_2$$

where for any input $Y \in \mathbb{R}^{T \times d}$, we have parameters $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$, $b_1 \in \mathbb{R}^{d_{ff}}$, and $b_2 \in \mathbb{R}^d$. The intermediate dimension d_{ff} is typically chosen to be larger than d (in practice, often $d_{ff} = 4d$) to enable richer non-linear transformations.

The theoretical foundations in [Yun et al., 2020] establish two key properties that validate our transformer-based approach to text regression. First, they prove that transformers are universal approximators of continuous sequence-to-sequence functions. That is, for any continuous function $f : \mathbb{R}^{d \times T} \to \mathbb{R}^{d \times T}$ with compact support and any $\epsilon > 0$, there exists a transformer network g such that:

$$\left(\int \|f(X) - g(X)\|_p^p dX \right)^{1/p} \leq \epsilon$$

for any $1 \leq p < \infty$. This universal approximation property holds both for permutation equivariant functions and, when positional encodings are added, for arbitrary continuous sequence-to-sequence functions.

Second, they formalize the notion of contextual representations through what they term "contextual mappings". For any finite set of sequences $\mathcal{L}$, a contextual mapping $q : \mathcal{L} \to \mathbb{R}^{1\times T}$ must satisfy:

(i) For any sequence $L \in \mathcal{L}$, the T entries in $q(L)$ are all distinct

(ii) For any $L, L' \in \mathcal{L}$ with $L \neq L'$, all entries in $q(L)$ and $q(L')$ are distinct

These properties ensure that each token's representation uniquely encodes both its position and its full context within the sequence. In our setting, this means that the representation $z_t^{(l,i)}$ of token t in document i at layer l captures not just the token's meaning in isolation, but its relationship with all other tokens in the document.

This theoretical framework provides a justification for our methodological choice. When analyzing text data, we require representations that can capture complex semantic relationships and long-range dependencies while being invariant to specific lexical choices. The universal approximation property guarantees that transformers can learn such representations, while the contextual mapping property ensures that these representations will preserve the rich contextual information.

3.3 Text Regression

Suppose a researcher has data that has n documents s_i in which every document, denoted as $\{s_i\}_{i=1}^n$, where each document s_i is associated with a corresponding label y_i. These labels do not need to be external to the text data, it can be derived from the text itself with great confidence. There also might be some document specific data x_i that captures for example author characteristics, or some treatment effect. We propose the following method when the researcher is interested in the association of text and the label.

We first use embeddings based on the transformer architecture to represent the document in the lower dimensional embedding space and then fit a second stage model to learn the association between the label of the document and the information within the text data.

Our first stage encoding leverages the power of transformer-based models, as described in the methodological overview. The ability of transformers to model token interactions and capture contextual information allows us to generate high-quality embeddings that preserve the semantic richness of the forum posts. Hence, the method we propose has two stages:

First stage is to get context dependent embeddings of documents.

$$e_i = g(s_i) + \epsilon_i \qquad (1)$$

After obtaining these latent representations, one can employ any suitable econometric model to estimate:

$$y_i = f(\beta e_i + \gamma x_i) + \varepsilon_i \qquad (2)$$

where $f()$ is an appropriate link function, β represents the coefficients for the latent text representations, γ represents the coefficients for the observation-specific covariates, and ε_i is the error term.

A potential concern in any two-stage procedure arises if the error terms from the first stage are systematically correlated with the error term in the second-stage regression. In classical econometric settings, such correlation may render the second-stage coefficient estimates biased and inconsistent if one attempts to interpret them as structural or causal parameters. However, our empirical framework does not seek a causal interpretation of these estimates. Instead, our approach is closer to a reduced-form or predictive regression, wherein the objective is to glean how textual content differs when referring to male vs. female subjects rather than to isolate any structural "causal effect" of specific linguistic features. In that spirit, We explicitly remove direct markers- pronouns, names, or family-member words—so that the remaining signals of gender orientation in the text cannot simply hinge on superficial tokens/words. This choice does not itself guarantee the first-stage errors are uncorrelated with the second-stage regression errors, but it does ensure that predictions are driven by subtler semantic patterns rather than by trivial keywords or obvious references. If our second-stage can still discriminate reliably after such redactions, it implies that distinctive linguistic dif-

ferences remain—even in the absence of explicit gender labels.

We also use embedding models that are estimated at or near a "population level". This substantially reduces sampling variability in the embeddings and offers more robust textual representations. In principle, if the embedding model captured all relevant semantic information perfectly, then any first-stage "errors" would be negligible. In practice, however, a perfectly universal embedding is rarely attainable, and any omitted textual features that matter for the second-stage outcome can, in principle, create correlation between the residuals. Thus, while population-level training does help mitigate noise, it does not eliminate the possibility that certain omitted features remain correlated with the second-stage residual.

With that being said, the two-stage regression framework is flexible enough to support causal analyses under the right experimental or quasi-experimental designs. For instance, in behavioral economics applications, one could randomly assign different message styles (e.g., supportive vs. formal) to participants and measure downstream effects like vaccination uptake, using embeddings to quantify subtle text differences. In marketing or platform design, an A/B test that randomly assigns landing-page text (short/casual vs. long/technical) could leverage embeddings to link linguistic variation to conversion rates. For policy or legal documents, quasi-random variation in drafting styles can generate exogenous shifts in text complexity or tone, allowing researchers to test whether simpler language fosters higher compliance. Finally, an instrumental-variables approach can treat random editorial decisions (or other exogenous factors) as valid instruments for textual content, so long as the chosen instrument meaningfully shifts the text's features without directly affecting outcomes. In all these scenarios, the embedding merely encodes the manipulated language attributes, and those become the causal "treatment"—demonstrating how a two-stage text regression can be adapted for causal inference when design conditions are met. In other words, the embedding is not itself the fundamental cause – the random assignment of a particular style is – but the embedding captures those stylistic or semantic differences, which is what the regression then uses to estimate a causal effect on the outcome.

3.3.1 First Stage Models

Transformer architectures can be broadly categorized into three types: encoder-only, decoder-only, and encoder-decoder models. Understanding the distinctions between these architectures is crucial for selecting the appropriate model for specific tasks, such as text regression.

Encoder-only transformers, exemplified by [Devlin et al., 2018] (Bidirectional Encoder Representations from Transformers, BERT), process the entire input sequence simultaneously. The encoder stack comprises multiple layers, each containing a self-attention mechanism and a feed-forward neural network. These models are characterized by their ability to consider bidirectional context, processing both left and right contexts for each token. This parallel processing allows for efficient computation, as all tokens are processed concurrently. Encoder-only transformers produce fixed-length representations for each input token, making them particularly well-suited for tasks that require a deep understanding of the input text, such as text classification, named entity recognition, and sentiment analysis.

Decoder-only transformers, such as those introduced by [Radford et al., 2019] (Generative Pre-trained Transformer, GPT2) or [Brown et al., 2020] (Generative Pre-trained Transformer, GPT) generate sequences in an autoregressive manner, producing one token at a time conditioned on the previously generated tokens. Unlike their encoder-only counterparts, decoder-only models are inherently unidirectional, leveraging past tokens in the sequence while masking future tokens to ensure causality during the generation process. Each layer in the decoder architecture incorporates masked self-attention, preventing the model from attending to tokens that have not yet been predicted. This sequential, autoregressive approach makes decoder-only transformers particularly well-suited for generative tasks such as language modeling, dialogue generation, and text completion. Their capacity to model long-range dependencies and generate contextually coherent outputs positions them as a powerful tool for a wide range of natural language generation applications.

Encoder-decoder transformers, such as [Raffel et al., 2019] (Text-to-Text Transfer Transformer, T5), consist of an encoder stack followed by a decoder stack. This architecture

is designed for sequence-to-sequence tasks and exhibits several key features. The decoder generates output tokens sequentially, employing a cross-attention mechanism that allows it to focus on relevant parts of the input by attending to the encoder's output. Additionally, encoder-decoder transformers possess an autoregressive property, where each generated token depends on previously generated tokens. These characteristics make encoder-decoder models particularly effective for tasks like machine translation, text summarization, and question answering.

For the text regression task proposed in this study, which aims to quantify language differences among different groups in an online forum, an encoder-only transformer is particularly appropriate. The primary objective is to obtain a fixed-length representation of the input text for subsequent analysis, which aligns well with the capabilities of encoder-only transformers. The bidirectional context and parallel processing features of these models allow for efficient and effective encoding of textual information from forum posts into a format amenable to regression analysis. This approach enables a nuanced examination of language use across different demographic groups, potentially revealing insights into gender-based or other group-specific linguistic patterns in anonymous online discussions.

The initial stage of our proposed methodology involves encoding the textual data into a lower-dimensional representation. This crucial step allows for the efficient processing and analysis of complex linguistic information. We have explored several options for this encoding process, each with its own set of advantages and limitations.

One approach involves utilizing pre-trained models, such as BERT (Bidirectional Encoder Representations from Transformers) and SBERT. These models leverage large-scale pre-training and capture rich linguistic knowledge, often requiring minimal fine-tuning for specific tasks. However, they may be computationally intensive and potentially biased towards their pre-training domain. Another option is the use of API embedding services. These services offer ease of implementation, regular updates, and often state-of-the-art performance, however the rate limits imposed by the companies that offer such services makes the process slow. Additionally, they may raise potential privacy concerns, create dependency on third-

party services, and incur associated costs. Sentence embeddings present a lightweight and fast alternative, particularly suitable for tasks requiring sentence-level representations. While efficient, this approach may sacrifice fine-grained token-level information, which could be crucial for certain analyses. For more specialized applications, custom transformer training can be employed. This approach allows for tailoring the model to the specific domain and task, potentially capturing unique patterns in the data. However, it requires significant computational resources and large amounts of domain-specific data. It is worth noting that the embedding dimensions of various models play a crucial role in their performance and computational efficiency. In this study, we use OpenAI's Ada models which utilize a 1536-dimensional embedding space, while BERT employs 768 dimensions, and SBERT operates with 386 dimensions. These differences in dimensionality can significantly impact the second stage performance of downstream tasks.

The selection of the most appropriate first stage model depends on various factors, including the specific requirements of the task, available computational resources, and the nature of the textual data under investigation. Researchers must carefully weigh these considerations to determine the most suitable encoding method for their particular study. Ultimately, the choice of encoding technique plays an important role in the overall effectiveness and efficiency of the two-stage text regression methodology.

Extracting embeddings from BERT and Sentence Embedding models

The pretraining process of BERT is crucial to its effectiveness and involves two main tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, approximately 15% of input tokens are masked, and the model is trained to predict these masked tokens based on the surrounding context. This task enables BERT to capture bidirectional context effectively. The NSP task involves predicting whether two sentences appear consecutively in the original text, helping the model understand sentence relationships.

During pretraining, BERT processes input text as a sequence of tokens, with special tokens added for specific purposes. The [CLS] token, introduced at the beginning of each

input sequence, serves a unique role. Initially, it acts as a placeholder, but through the self-attention mechanisms in BERT's layers, it accumulates information from the entire input sequence. This makes the [CLS] token particularly useful for tasks that require a single, fixed-length representation of the entire input.

When using BERT for embedding generation, researchers have several options. One common approach is to use the [CLS] token embedding from the final layer of BERT. Mathematically, if we denote the final hidden state of the [CLS] token as h_{CLS}, the embedding can be represented as:

$$e_{CLS} = W * h_{CLS} + b$$

where W is a weight matrix and b is a bias term, which can be fine-tuned for specific tasks.

Another option is to use the mean of all token embeddings from the final layer. This approach can be represented as:

$$e_{mean} = (1/n) * \sum(h_i)$$

where h_i represents the final hidden state of the i-th token, and n is the number of tokens.

Both methods have their merits. The [CLS] token embedding is often preferred for sentence-level tasks as it theoretically captures the essence of the entire input. However, taking the mean of all token embeddings can sometimes provide a more comprehensive representation, especially for longer texts or when fine-grained token-level information is crucial.

Researchers can also consider using embeddings from intermediate layers of BERT, as these may capture different levels of linguistic abstraction. Some studies have shown that combining embeddings from multiple layers can lead to improved performance on certain tasks.

Moving to sentence embedding models, these offer an alternative approach focused on generating fixed-length representations for entire sentences or short texts. Models like Universal

Sentence Encoder (USE) or Sentence-BERT are specifically designed for this purpose. These models often employ techniques like siamese or triplet network structures during training to ensure that semantically similar sentences have similar embeddings in the vector space.

Sentence embedding models typically produce a single vector representation for an entire input text, regardless of its length. This approach can be particularly advantageous when dealing with variable-length inputs or when computational efficiency is a priority. The mathematical representation of a sentence embedding can be simplified as:

$$e_{sentence} = f(s)$$

where f is the sentence embedding function, and s is the input sentence.

One key advantage of sentence embedding models is their ability to capture semantic similarity between sentences more effectively than token-level models. This makes them particularly useful for tasks such as semantic search, clustering, or measuring text similarity.

3.3.2 Second-Stage Models

After obtaining the latent representations $e_i \in \mathbb{R}^d$ of the text data through dimensionality reduction, we proceed to estimate the relationship between these representations and the outcome variable $y_i \in \mathbb{R}$ using three different models: Logistic Regression (Logit), Random Forests, and Multi-Layer Perceptrons (MLP). Below, we provide the mathematical formulations, objective functions, and optimization problems for each model.

Logistic Regression

Logistic regression is used to model binary outcome variables $y_i \in \{0, 1\}$. Given the predictors $e_i \in \mathbb{R}^d$ (latent text representations) and $x_i \in \mathbb{R}^p$ (observation-specific covariates), the logistic regression model estimates the probability $p_i = P(y_i = 1|e_i, x_i)$ using the logistic function:

$$p_i = \frac{1}{1 + \exp(-(\beta^\top e_i + \gamma^\top x_i))},$$

where $\beta \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}^p$ are the coefficients to be estimated.

The objective function for logistic regression is the negative log-likelihood of the observed data:

$$L(\beta, \gamma) = -\sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right].$$

The optimization problem is:

$$\min_{\beta \in \mathbb{R}^d, \gamma \in \mathbb{R}^p} L(\beta, \gamma).$$

Regularization techniques like L1 (LASSO) or L2 (Ridge) can be added to penalize large coefficients, leading to:

$$\min_{\beta, \gamma} L(\beta, \gamma) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\gamma\|_2^2,$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters.

Random Forests

Random forests are ensemble learning methods that aggregate the predictions of multiple decision trees to improve robustness and accuracy. Let $\{T_b\}_{b=1}^B$ denote B decision trees trained on bootstrap samples of the data. Each tree T_b predicts the outcome $\hat{y}_i^{(b)}$ for observation i.

The prediction of the random forest is the average of the predictions of all trees:

$$\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_i^{(b)}.$$

Each tree T_b is constructed by solving the following recursive optimization problem for node splitting:

$$\min_{k,c} \left(\sum_{i \in S_L(k,c)} (y_i - \bar{y}_L)^2 + \sum_{i \in S_R(k,c)} (y_i - \bar{y}_R)^2 \right),$$

where $S_L(k, c) = \{i : X_{ik} \leq c\}$ and $S_R(k, c) = \{i : X_{ik} > c\}$ are the left and right child nodes resulting from a split on the k-th covariate at threshold c, and $\bar{y}_L$ and $\bar{y}_R$ are the mean outcomes for observations in $S_L(k, c)$ and $S_R(k, c)$, respectively.

Random forests do not explicitly optimize at the ensemble level, but they implicitly aim to minimize prediction error by combining multiple low-bias, high-variance models.

Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron (MLP) is a type of feedforward neural network consisting of multiple layers of nodes (neurons), including an input layer, one or more hidden layers, and an output layer. Each layer performs a linear transformation followed by a non-linear activation function.

For a given observation i, the MLP model is defined as:

$$\hat{y}_i = f\left(W^{(L)}\phi\left(W^{(L-1)}\phi(\cdots\phi(W^{(1)}z_i + b^{(1)})\cdots) + b^{(L-1)}\right) + b^{(L)}\right),$$

where $z_i = [e_i; x_i] \in \mathbb{R}^{d+p}$ is the concatenated input vector. $W^{(l)} \in \mathbb{R}^{h_l \times h_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{h_l}$ are the weight matrix and bias vector for layer l. $\phi(\cdot)$ is a non-linear activation function (e.g., ReLU: $\phi(x) = \max(0, x)$). $f(\cdot)$ is the output activation function, which could be a sigmoid function for binary classification.

The objective function for training an MLP is the empirical risk minimization of a loss function $\ell(\cdot, \cdot)$ over the training data:

$$\min_{\{W^{(l)}, b^{(l)}\}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \hat{y}_i),$$

where $\ell(y_i, \hat{y}_i) = -\left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\right)$ is the cross-entropy loss for binary classification.

The optimization is typically performed using stochastic gradient descent (SGD) or its variants, with regularization terms (e.g., weight decay or L2 regularization) added to the loss function.

3.4 Data

Our analysis utilizes data from the Economics Job Market Rumors (EJMR) forum, an anonymous online platform for economists. The dataset, originally compiled by Wu2018, comprises 1,048,574 observations from threads containing at least one gendered post, as identified by Wu's methodology. Note that this implies that not every post has gendered language in them. This substantial corpus provides a rich source of textual data for analyzing language differences among various groups in the economics community. Importantly, our analysis does not predict the gender of the author, but rather whether the text is about or refers to a female or male. This distinction is crucial for understanding the nature of our study and its implications.

The dataset includes 444,810 gendered posts, representing approximately 42.4% of the total observations. These posts span 138,477 unique threads, offering a diverse range of discussions within the economics field. The temporal coverage of the data extends from October 2013 to January 2018, providing a comprehensive view of discourse on the EJMR forum over this period.

To illustrate the nature of the content in our dataset, we present a selection of posts in Table II. We caution readers that some of the language used in these posts may be considered offensive or inappropriate.

Table I: Examples of Posts from the EJMR Dataset

EJMR example posts
1." Yeah, but are any of them popular with female *** big *** healthy young undergrads? That's really all that matters in life. "
2." Gave up on women all people really got a dog last week I think this will turn out to be one of the best decisions I ever made. Should have done this a long time ago. Just me and Boomer now."
3." Baby Rudin is an excellent book for learning Analysis for the first time. Probably the most popular too. That's all there is to it. Also, are you trolling with "harder to learn from"? That's not a plus for a book! And, I have never read KF's analysis book, but I do have their functional analysis book, and it's awful. Maybe their RA is better though. "
4. "For this type of job, you're not being hired based on your perceived knowledge set, but instead based on your perceived IQ and analytic reasoning skills."
5. "Successfully went from aspie to social alpha. Two critical points: 1) How you perceive yourself sets the tone for how others will treat you. Dress well, groom yourself well. Look into the mirror. Even if you're fat, dress down your flaws (dark colours, layer, focus on fit) 2) There's a two-way causality... How others treat you also affects how you perceive yourself. This means you need to spend time with people who treat you well. Eliminate people who treat you poorly from your social circle. Also, you might feel uncomfortable at times, like say, near a hot cashier or something. In low stakes situations (you won't see the person ever again), it might be worthwhile to push your limits. Smile, say hi to the cashier without looking away. Say yes to social events, gatherings, and interactions. People aren't as evil as you think. Sometimes, admitting that you're aspie in a big group will make them behave in a more friendly and accommodative manner. Cheers."
6. "No, but I have heard of the phrase "meth mouth". "
7. "Chicago bros don't really need to promote their school... "
8. "I'm 30 kilos overweight and addicted to alcohol and painkillers. I'm also stuck on antidepressants and I have no friends or girlfriend."
9." I don't know *** well but have had a number of informal meetings with him and he has always been friendly, helpful, and engaging. I find the rumors on this board of fits of rage etc. very difficult to reconcile with what I have seen. "
10. "Possible downside of marrying a lawyer: Is she the vindictive type? Will she sue the pants off you or kill you in custody proceedings if things go wrong in the marriage? If not, then go ahead. If yes, then... at least sign a prenup. "

A crucial distinction in our approach lies in the labeling strategy employed. Unlike

previous studies that relied on predefined lists of gender classifiers, our methodology utilizes Large Language Models (LLMs) to determine the gender orientation of each post and then removing explicit gender identifiers to reduce bias in the prediction stage. This approach offers several advantages over traditional methods. By removing explicit gender identifiers (e.g., names, pronouns, family member identifiers) before classification, we mitigate the risk of misclassification based on superficial textual features as well as predicting the gender by looking only on these identifiers. Our method allows for gender prediction based on the semantics and context of the entire post, rather than relying on the presence of specific words or phrases. This semantic focus potentially yields more accurate classifications compared to dictionary-based approaches, especially for posts with subtle or implicit gender references. Furthermore, the AI-based labeling strategy can be adapted to categorize posts along various dimensions beyond gender, such as academic subfields or career stages, by modifying the prompt used in the classification process.

The labeling process involves a two-step procedure that enhances the robustness of our analysis. First, an AI model classifies each post as "FEMALE," "MALE," or "NEUTRAL" based on the content and context of the text. This initial classification leverages the advanced language understanding capabilities of modern AI models to capture nuanced gender references. Subsequently, we employ an identifier removal step, where gender-specific words and names are replaced with neutral placeholders (e.g., "[PRONOUN]," "[FAMILY_MEMBER]," "[NAME]") while preserving the overall sentence structure. This crucial step ensures that our analysis focuses on the underlying semantics of the posts rather than relying on explicit gender markers. By combining these two steps, we create a dataset that allows for a more nuanced exploration of language differences based on gender, while simultaneously reducing the potential for bias introduced by superficial textual features.

This comprehensive approach enables us to analyze language differences based on the underlying semantics of the posts, rather than relying on explicit gender markers. Moreover, the flexibility of our labeling strategy allows for the exploration of various group dynamics within the economics community, extending beyond gender to potentially include other de-

mographic or professional categories of interest. By leveraging advanced AI technologies and employing a thoughtful, multi-step labeling process, we aim to provide a more accurate and insightful analysis of the discourse within the economics community as represented on the EJMR forum.

Table II: Examples of masking and labeling from the EJMR Dataset

EJMR example masked posts	*gender*
1."Yeah, but are any of them popular with [GENDER] big healthy young undergrads? That's really all that matters in life."	FEMALE
2."Gave up on [GENDER] all people really got a dog last week I think this will turn out to be one of the best decisions I ever made. Should have done this a long time ago. Just me and Boomer now "	MALE
3."Baby Rudin is an excellent book for learning Analysis for the first time. Probably the most popular too. That's all there is to it. Also, are you trolling with ẗharder to learn from?̈ That's not a plus for a book! And, I have never read KF's analysis book, but I do have their functional analysis book, and it's awful. Maybe their RA is better though."	NEUTRAL
4."For this type of job, you're not being hired based on your perceived knowledge set, but instead based on your perceived IQ and analytic reasoning skills"	NEUTRAL
5. "Successfully went from aspie to social alpha. Two critical points: 1) How you perceive yourself sets the tone for how others will treat you. Dress well, groom yourself well. Look into the mirror. Even if you're fat, dress down your flaws (dark colours, layer, focus on fit) 2) There's a two-way causality... How others treat you also affects how you perceive yourself. This means you need to spend time with people who treat you well. Eliminate people who treat you poorly from your social circle. Also, you might feel uncomfortable at times, like say, near a hot cashier or something. In low stakes situations (you won't see the person ever again), it might be worthwhile to push your limits. Smile, say hi to the cashier without looking away. Say yes to social events, gatherings, and interactions. People aren't as evil as you think. Sometimes, admitting that you're aspie in a big group will make them behave in a more friendly and accommodative manner. Cheers"	MALE
6."No, but I have heard of the phrase meth mouth"	NEUTRAL
7. "Chicago [GENDER] don't really need to promote their school..."	MALE
8. "I'm 30 kilos overweight and addicted to alcohol and painkillers. I'm also stuck on antidepressants and I have no friends or [PARTNER]."	MALE
9. "I don't know [NAME] well but have had a number of informal meetings with [PRONOUN] and [PRONOUN] has always been friendly, helpful, and engaging. I find the rumors on this board of fits of rage etc. very difficult to reconcile with what I have seen"	MALE
10. "Possible downside of marrying a lawyer: Is [PRONOUN] the vindictive type? Will [PRONOUN] sue the pants off you or kill you in custody proceedings if things go wrong in the marriage? If not, then go ahead. If yes, then... at least sign a prenup"	FEMALE

3.5 Results

In this section, we present the results of our two-stage text regression methodology applied to the EJMR dataset. We explore various combinations of first stage text encoding models and second-stage classification models to evaluate their effectiveness in predicting the gender orientation of forum posts.

3.5.1 Evaluation Metrics

For each combination of first stage and second-stage models, we report a comprehensive set of evaluation metrics to assess performance. Accuracy is calculated as the proportion of correctly classified instances out of the total number of instances, providing an overall measure of correctness. Precision is determined by the ratio of true positive predictions to the total number of predicted positives, thereby evaluating the quality of positive predictions. Recall (Sensitivity) measures the ratio of true positive predictions to all actual positive instances, indicating the model's ability to identify positive cases effectively. To balance precision and recall, we report the F1-Score, which represents the harmonic mean of these two metrics.

In addition to these metrics, we include Macro Avg and Weighted Avg scores. Macro Avg computes the unweighted mean of precision, recall, and F1-score across all classes, treating each class equally. This metric is useful for evaluating performance when dealing with class imbalance, as it does not take the size of each class into account. Weighted Avg, on the other hand, computes the mean of precision, recall, and F1-score, but weights each class's contribution according to its size. This allows the metric to reflect the overall performance of the model, especially in cases where some classes dominate the dataset. Together, these metrics provide a robust framework for evaluating and comparing the performance of different model combinations in capturing the association between text data and labels.

3.5.2 Numerical Results

Tables III through VI present the classification metrics for various combinations of first stage text encoding models and second-stage classification models applied to the EJMR dataset. These tables showcase the performance of different model combinations in predicting the gender orientation of forum posts. Note that we kept the hyperparameters of the second-stage models fixed across all experiments, as detailed in Tables V and VI. This approach ensures a fair comparison between different embedding methods and highlights the robustness of our results.

Our two-stage text regression methodology demonstrates robust performance in predicting gender of the posts across various combinations of first stage text encoding models and second-stage classification models. This strong predictability is particularly noteworthy given the absence of explicit gender identifiers in the text, suggesting that subtle linguistic patterns are effectively captured by our approach.

Comparing the first stage encoding models, we observe that the OpenAI-Ada-v3-small embeddings consistently outperform other encoding methods across all second-stage classifiers. For instance, the Multi-Layer Perceptron (MLP) achieves the highest accuracy of 0.79 when using OpenAI-Ada-v3-small embeddings, compared to 0.76, 0.75, and 0.75 for BERT-MEAN, SBERT, and BERT-CLS embeddings, respectively. This superior performance of OpenAI-Ada-v3-small embeddings is consistent across all evaluation metrics, including precision, recall, and F1-score.

Among the second-stage models, the MLP classifier demonstrates the strongest performance across all embedding types, followed closely by XGBoost. The Lasso Logistic regression consistently shows the weakest performance, suggesting that the relationship between text embeddings and gender orientation may be non-linear or require more complex feature interactions.

Interestingly, even the simplest models in our framework, such as Lasso Logistic regression, achieve accuracy above 70% across all embedding types. This observation supports the

interpretation of our procedure as a form of model distillation, where the complex patterns learned by sophisticated language models in the first stage are effectively compressed into a format that allows much smaller models to predict gender orientation with considerable accuracy.

Our methodology can be interpreted as a distillation approach introducedn in the seminal paper, [Hinton et al., 2015], where we initially use a powerful and computationally expensive Large Language Model (LLM) to label the posts. This process allows us to capture nuanced linguistic patterns and contextual information. Subsequently, we utilize much smaller and more cost-effective models to classify the posts based on these labels. This two-stage approach enables us to leverage the strengths of advanced LLMs while maintaining computational efficiency and reducing resource requirements in the classification stage.

It's important to note that the performance metrics presented in these tables represent a conservative estimate of the model's capabilities. If we were to include more data or retain gender identifiers in the text, we would likely observe improved performance across all model combinations. The current results demonstrate the effectiveness of our approach even with a constrained dataset, highlighting the potential for even greater accuracy in other applications where such limitations may not apply.

Table III: Comparison of Classification Metrics for BERT MEAN Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
Random Forest *hyperparameters: Table V*	FEMALE	0.75	0.77	0.76
	MALE	0.73	0.71	0.72
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
LightGBM *hyperparameters: Table V*	FEMALE	0.77	0.74	0.75
	MALE	0.71	0.74	0.73
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
XGBoost *hyperparameters: Table V*	FEMALE	0.77	0.75	0.76
	MALE	0.72	0.74	0.73
	Accuracy		0.75	
	Macro Avg	0.75	0.75	0.75
	Weighted Avg	0.75	0.75	0.75
MLP *hyperparameters: Table VI*	FEMALE	0.80	0.74	0.77
	MALE	0.73	0.79	0.76
	Accuracy		0.76	
	Macro Avg	0.76	0.76	0.76
	Weighted Avg	0.77	0.76	0.76
Lasso Logistic $\lambda = 0.001$	FEMALE	0.74	0.71	0.73
	MALE	0.68	0.71	0.70
	Accuracy		0.71	
	Macro Avg	0.71	0.71	0.71
	Weighted Avg	0.71	0.71	0.71

Table IV: Comparison of Classification Metrics for SBERT embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
Random Forest *hyperparameters: Table V*	FEMALE	0.76	0.75	0.75
	MALE	0.72	0.72	0.72
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
LightGBM *hyperparameters: Table V*	FEMALE	0.76	0.74	0.75
	MALE	0.71	0.74	0.72
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
XGBoost *hyperparameters: Table V*	FEMALE	0.77	0.74	0.76
	MALE	0.72	0.74	0.73
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
Multi Layer Perceptron *hyperparameters: Table VI*	FEMALE	0.79	0.73	0.76
	MALE	0.72	0.78	0.75
	Accuracy		0.75	
	Macro Avg	0.75	0.76	0.75
	Weighted Avg	0.76	0.75	0.75
Lasso Logistic $\lambda = 0.001$	FEMALE	0.74	0.71	0.72
	MALE	0.68	0.71	0.69
	Accuracy		0.71	
	Macro Avg	0.71	0.71	0.71
	Weighted Avg	0.71	0.71	0.71

Table V: Comparison of Classification Metrics for BERT-[CLS] Token Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
Random Forest *hyperparameters: Table V*	FEMALE	0.71	0.78	0.74
	MALE	0.71	0.63	0.67
	Accuracy		0.71	
	Macro Avg	0.71	0.70	0.70
	Weighted Avg	0.71	0.71	0.71
LightGBM *hyperparameters: Table V*	FEMALE	0.75	0.75	0.75
	MALE	0.71	0.71	0.71
	Accuracy		0.73	
	Macro Avg	0.73	0.73	0.73
	Weighted Avg	0.73	0.73	0.73
XGBoost *hyperparameters: Table V*	FEMALE	0.75	0.76	0.76
	MALE	0.72	0.71	0.72
	Accuracy		0.74	
	Macro Avg	0.74	0.74	0.74
	Weighted Avg	0.74	0.74	0.74
MLP *hyperparameters: Table VI*	FEMALE	0.79	0.73	0.76
	MALE	0.71	0.77	0.74
	Accuracy		0.75	
	Macro Avg	0.75	0.75	0.75
	Weighted Avg	0.75	0.75	0.75
Lasso Logistic $\lambda = 0.001$	FEMALE	0.74	0.72	0.73
	MALE	0.69	0.71	0.70
	Accuracy		0.71	
	Macro Avg	0.71	0.71	0.71
	Weighted Avg	0.72	0.71	0.71

Table VI: Comparison of Classification Metrics for OpenAI-Ada-v3-small embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
Random Forest *hyperparameters: Table V*	FEMALE	0.77	0.78	0.78
	MALE	0.75	0.73	0.74
	Accuracy		0.76	
	Macro Avg	0.76	0.76	0.76
	Weighted Avg	0.76	0.76	0.76
LightGBM *hyperparameters: Table V*	FEMALE	0.79	0.75	0.77
	MALE	0.73	0.77	0.75
	Accuracy		0.76	
	Macro Avg	0.76	0.76	0.76
	Weighted Avg	0.76	0.76	0.76
XGBoost *hyperparameters: Table V*	FEMALE	0.80	0.77	0.78
	MALE	0.75	0.78	0.76
	Accuracy		0.77	
	Macro Avg	0.77	0.77	0.77
	Weighted Avg	0.78	0.77	0.77
Multi Layer Perceptron *hyperparameters: Table VI*	FEMALE	0.84	0.76	0.80
	MALE	0.75	0.83	0.79
	Accuracy		0.79	
	Macro Avg	0.79	0.79	0.79
	Weighted Avg	0.80	0.79	0.79
Lasso Logistic $\lambda = 0.001$	FEMALE	0.75	0.73	0.74
	MALE	0.69	0.72	0.70
	Accuracy		0.72	
	Macro Avg	0.72	0.72	0.72
	Weighted Avg	0.72	0.72	0.72

3.5.3 Out-of-Sample Analysis of Gender Predictions

To gain deeper insights into the linguistic patterns associated with gender on the EJMR forum, we conducted an out-of-sample analysis of the posts most strongly predicted as male

or female by our best-performing model. This analysis focuses on the top 500 posts with the highest prediction probabilities for each gender, allowing us to examine the most distinctive language features that our model associates the language and a post being about a male or female.

We employed two complementary visualization techniques to explore these high-confidence predictions: word clouds and Latent Dirichlet Allocation (LDA)[3] topic modeling. Word clouds provide an intuitive representation of word frequency, with the size of each word proportional to its occurrence in the corpus. This allows for a quick visual assessment of the most prominent terms associated with each gender. LDA topic modeling, on the other hand, uncovers themes within the text data, offering a more nuanced view of the underlying topics.

Figure 3.1 presents the word cloud for posts predicted as female, while Figure 3.2 shows the word cloud for male posts. The LDA topic modeling results for both genders are summarized in Table VII.

The female word cloud in Figure 3.1 reveals a prominence of terms related to family and personal relationships, such as "family," "member," "son," "partner," and "children." This suggests that posts predicted as female tend to discuss personal and familial matters more frequently. Additionally, words like "work," "life," and "time" indicate a focus on work-life balance issues.

In contrast, the male word cloud in Figure 3.2 is dominated by terms related to academic achievement and professional evaluation, such as "great," "good," "economist," "scholar," and "influential." Words like "methodological," "theory," and "evidence" suggest a stronger emphasis on research and academic discourse. The prominence of comparative terms like "ahead" and "outstanding" may indicate a tendency towards more competitive or evaluative language in male-predicted posts.

---

[3]We go over LDA in the Appendix

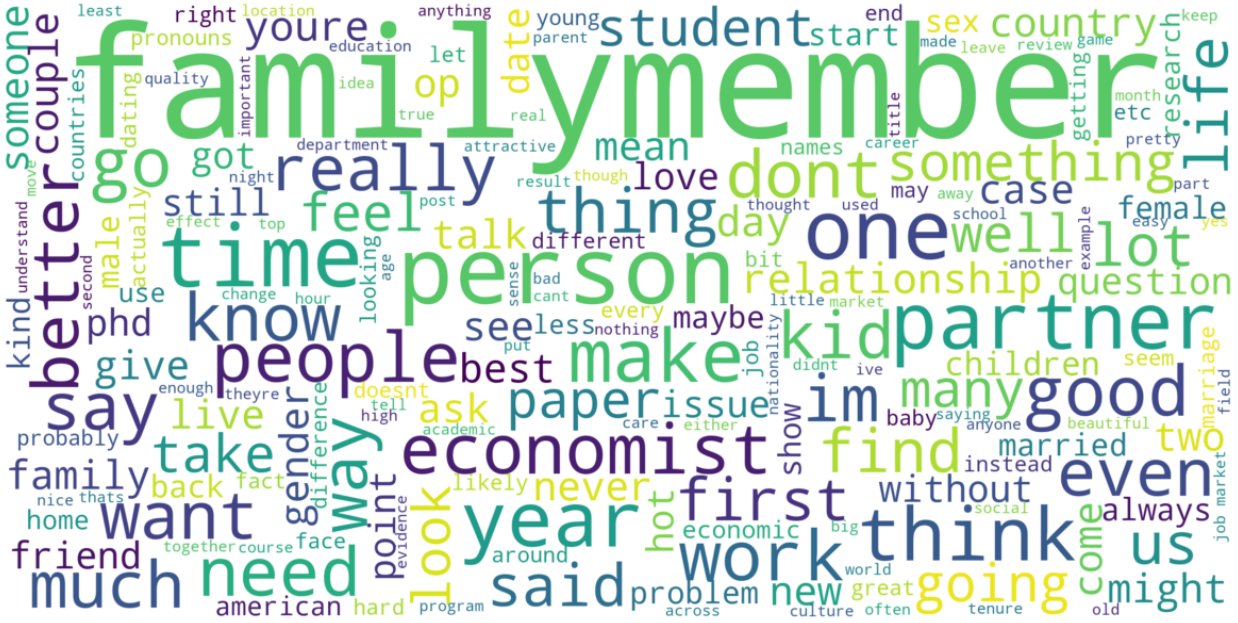


Figure 3.1: The most frequent words associated with predicting the FEMALE gender in out-of-sample posts. The analysis is based on the top 500 posts with the highest predicted probabilities for the FEMALE category.

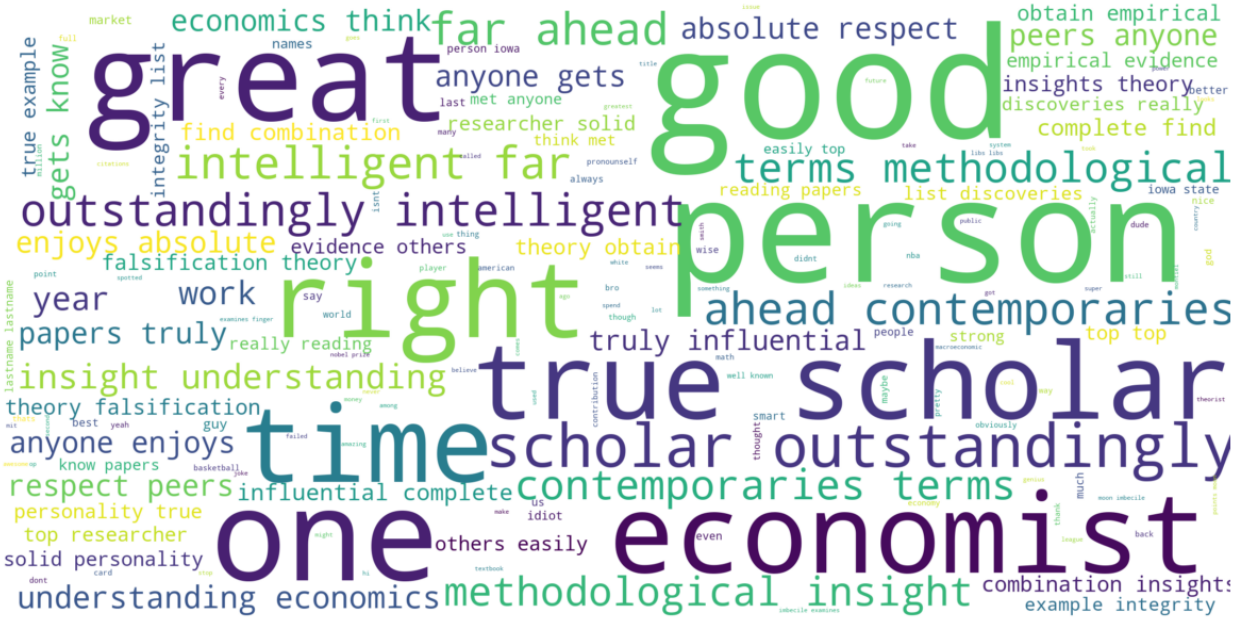


Figure 3.2: The most frequent words associated with predicting the MALE gender in out-of-sample posts. The analysis is based on the top 500 posts with the highest predicted probabilities for the MALE category.

Table VII: LDA Topics for Males and Females (Top 5 words per topic)

Gender	LDA Topics
Males	Topic 1: pronoun, theory, true, papers, far Topic 2: pronoun, like, right, last_name, good Topic 3: person, strong, state, great, iowa Topic 4: pronoun, good, http, work, economic Topic 5: nice, obviously, guy, smart, montiel
Females	Topic 1: pronoun, family_member, just, like, person Topic 2: pronoun, family_member, relationship, didn, good Topic 3: family_member, data, pronoun, like, people Topic 4: pronoun, person, 00, market, university Topic 5: family_member, person, pronoun, like, new

The LDA topic modeling results (Table VII) provide further nuance to these observations. Female-associated topics show a mix of personal relationships (Topics 1 and 2), professional

concerns (Topics 3 and 4), and general discussion (Topic 5). The recurring presence of "family_member" across multiple topics underscores the importance of family-related discussions in female-predicted posts.

Male-associated topics, on the other hand, appear more focused on academic and professional matters. Topics 1 and 2 emphasize theoretical work and academic discourse, while Topics 3 and 4 suggest discussions about professional achievements and economic concepts. The presence of location-specific terms (e.g., "iowa" in Topic 3) might indicate discussions about job markets or specific institutions.

These findings align with and extend previous research on gender differences in online communication, such as the work by [Wu, 2018]. Our analysis suggests that in an anonymous forum setting, there are discernible differences in the language and topics associated with male and female genders. Female-predicted posts tend to blend personal and familial concerns more frequently, while male-predicted posts focus more heavily on academic achievements and theoretical discussions.

It's important to note that these patterns reflect the model's predictions based on learned associations, rather than inherent gender differences. They probably are influenced by societal expectations, forum dynamics. Furthermore, this analysis focuses on the extremes of the prediction spectrum and may not be representative of all posts on the forum.

3.6 Conclusion

This paper introduces a novel two-stage text regression methodology that leverages transformer-based encodings to enhance the integration of textual data in econometric models. Our approach addresses the limitations of traditional bag-of-words methods by capturing richer semantic information and contextual nuances. The first stage employs advanced natural language processing techniques to represent textual data in a lower-dimensional space, while the second stage estimates the association between the outcome variable and the encoded text information.

Our methodology also offers a powerful tool for understanding association preferences in

various contexts. For example, in the field of marketing, researchers could use this approach to analyze customer reviews and identify the linguistic features most strongly associated with positive or negative product experiences. This could provide valuable insights for product development and customer service strategies.

Empirically, our results demonstrate the effectiveness of this two-stage approach in capturing gender-associated language patterns on the Economics Job Market Rumors forum. The best-performing model combination (OpenAI-Ada-v3-small embeddings with a Multi-Layer Perceptron classifier) achieved a balanced accuracy of 0.79 in predicting the gender orientation of posts, even after removing explicit gender identifiers. This high predictability suggests that subtle linguistic patterns are effectively captured by our approach.

The out-of-sample analysis revealed distinct themes associated with male and female-predicted posts. Female-associated language tended to blend personal and familial concerns more frequently, while male-associated language focused more heavily on academic achievements and theoretical discussions. These findings align with and extend previous research on gender differences in online communication.

Our approach offers several strengths. First, it provides a flexible framework that can be adapted to various research questions and domains. Second, by leveraging advanced NLP techniques, it captures nuanced linguistic patterns that might be missed by simpler methods. Third, the two-stage structure allows for efficient processing of large text datasets while maintaining interpretability in the second stage.

However, there are also limitations to consider. The reliance on pre-trained language models in the first stage may introduce biases present in the training data of these models. Additionally, while our method captures complex linguistic patterns, it may not fully account for the context-dependent nature of language use in specific domains.

The interdisciplinary nature of this approach, combining techniques from natural language processing, machine learning, and econometrics, opens up exciting possibilities for cross-pollination of ideas between these fields. Future research could explore domain-specific pre-training of language models, incorporate more contextual information into the analysis,

and investigate the application of this method to other forms of unstructured data, such as images or audio.

The potential impact of this methodology extends beyond academic research. By providing a robust framework for analyzing large-scale textual data, it could inform policy-making decisions in areas such as social media regulation, workplace diversity initiatives, or public health communication strategies. In the business world, it could enhance market research techniques, improve customer relationship management, and provide deeper insights into brand perception and consumer behavior.

BIBLIOGRAPHY

[Abadie et al., 2022] Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2022). When Should You Adjust Standard Errors for Clustering?*. *The Quarterly Journal of Economics*, 138(1):1–35.

[Arellano, 1987] Arellano, M. (1987). Practitioners' corner: Computing robust standard errors for within-groups estimators*. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

[Athey and Imbens, 2017] Athey, S. and Imbens, G. (2017). Chapter 3 - the econometrics of randomized experimentsa. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 73–140. North-Holland.

[Ba et al., 2016] Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

[Bajari et al., 2023] Bajari, P., Burdick, B., Imbens, G., Masoero, L., McQueen, J., Richardson, T., and Rosen, I. (2023). Experimental design in marketplaces. *Statistical Science*.

[Baker and Wurgler, 2006a] Baker, M. and Wurgler, J. (2006a). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4):1645–1680.

[Baker and Wurgler, 2006b] Baker, M. and Wurgler, J. (2006b). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680.

[Baker, 1994] Baker, S. G. (1994). The multinomial-poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(4):495–504.

[Bakshy et al., 2014] Bakshy, E., Eckles, D., and Bernstein, M. S. (2014). Designing and deploying online field experiments.

[Bana, 2021] Bana, S. H. (2021). Using language models to understand wage premia. *Institute for Human-Centered Artificial Intelligence*.

[Berman and Van den Bulte, 2022] Berman, R. and Van den Bulte, C. (2022). False discovery in a/b testing. *Manage. Sci.*, 68(9):6762–6782.

[Bettman, 1979] Bettman, J. R. (1979). Memory factors in consumer choice: A review. *Journal of Marketing*, 43(2):37–53.

[Birch, 1963] Birch, M. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(1):220–233.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Böhning, 1992] Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200.

[Böhning and Lindsay, 1988] Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663.

[Boskin, 1974] Boskin, M. J. (1974). A conditional logit model of occupational choice. *Journal of Political Economy*, 82(2, Part 1):389–398.

[Boyd et al., 2004] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press.

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

[Buchholz, 2021] Buchholz, N. (2021). Spatial equilibrium, search frictions, and dynamic efficiency in the taxi industry. *The Review of Economic Studies*, 89(2):556–591.

[Bühlmann and Van De Geer, 2011] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[Cai et al., 2015] Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

[Chelba et al., 2013] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

[Chen et al., 2021] Chen, Y., Han, B., and Pan, J. (2021). Sentiment trading and hedge fund returns. *Journal of Finance*, 76(4):2001–2033.

[Cohen, 1988] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition.

[Cohen et al., 2016] Cohen, P., Hahn, R. W., Hall, J., Levitt, S. D., and Metcalfe, R. (2016). Using big data to estimate consumer surplus: The case of uber. NBER Working Paper w22627, National Bureau of Economic Research. Available at SSRN: https://ssrn.com/abstract=2837639.

[Cooprider and Nassiri, 2023] Cooprider, J. and Nassiri, S. (2023). Science of price experimentation at amazon. In *AEA 2023, NABE 2023*.

[Crespi, 2019] Crespi, C. (2019). *Design and Analysis of Cluster Randomized Trials*.

[Davidson et al., 2010] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., and Livingston, B. (2010). The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pages 293–296.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Deng et al., 2013] Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 123–132, New York, NY, USA. Association for Computing Machinery.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dieudonné, 2011] Dieudonné, J. (2011). *Foundations of Modern Analysis*. Read Books Ltd.

[Dominitz and Sherman, 2005] Dominitz, J. and Sherman, R. P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory*, 21(4):838–863.

[Donaldson and Storeygard, 2016] Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98.

[Editor, 1986] Editor (1986). Hyphenation exception log. *TUGboat*, 7(3):145.

[Fagan and Iyengar, 2018] Fagan, F. and Iyengar, G. (2018). Unbiased scalable softmax optimization. *arXiv preprint arXiv:1803.08577*.

[Fan et al., 2015] Fan, Y., Pastorello, S., and Renault, E. (2015). Maximization by parts in extremum estimation. *The Econometrics Journal*, 18(2):147–171.

[Fithian et al., 2014] Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

[Frankel and Volij, 2011] Frankel, D. M. and Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory*, 146(1):1–38.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

[Fu and Knight, 2000] Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378.

[Furuya et al., 2024] Furuya, T., de Hoop, M. V., and Peyré, G. (2024). Transformers are universal in-context learners.

[Gentzkow et al., 2019a] Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–74.

[Gentzkow et al., 2019b] Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

[Geshkovski et al., 2024] Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. (2024). A mathematical perspective on transformers.

[Goldberg and Levy, 2014] Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

[Goossens et al., 1994] Goossens, M., Mittelbach, F., and Samarin, A. (1994). *The LaTeX Companion*. Addison-Wesley.

[Gopal and Yang, 2013] Gopal, S. and Yang, Y. (2013). Distributed training of large-scale logistic models. In *International Conference on Machine Learning*, pages 289–297. PMLR.

[Guo et al., 2021] Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. (2021). Machine learning for variance reduction in online experiments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8637–8648. Curran Associates, Inc.

[Hayes and Moulton, 2017] Hayes, R. and Moulton, L. (2017). *Cluster randomised trials, second edition*. CRC Press, United States.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Heiss, 2002] Heiss, F. (2002). Structural choice analysis with nested logit models. *The Stata Journal*, 2(3):227–252.

[Hesterberg and Knight, 2024] Hesterberg, T. and Knight, B. (2024). Power analysis for experiments with clustered data, ratio metrics, and regression for covariate adjustment.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.

[Horton, 2023] Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

[Imbens and Rubin, 2015] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[Jennrich, 1969] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

[Kelly et al., 2019] Kelly, B. T., Manela, A., and Moreira, A. (2019). Text selection. Working Paper 26517, National Bureau of Economic Research.

[Knuth, 1984] Knuth, D. E. (1984). *The TeX book*. Addison-Wesley.

[Knuth, 1986a] Knuth, D. E. (1986a). *TEX: The Program*. Addison-Wesley.

[Knuth, 1986b] Knuth, D. E. (1986b). *Computer Modern Typefaces*. Addison-Wesley.

[Knuth, 1986c] Knuth, D. E. (1986c). *The Metafont book*. Addison-Wesley.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

[Lamport, 1994] Lamport, L. (1994). *LATEX: A Document Preparation System*. Addison-Wesley, 2nd edition.

[Lang, 1996] Lang, J. B. (1996). On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):253–266.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

[Lee et al., 2016] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

[Markovic et al., 2017] Markovic, J., Xia, L., and Taylor, J. (2017). Unifying approach to selective inference with applications to cross-validation. *arXiv preprint arXiv:1703.06559*.

[Massey, 1951] Massey, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

[McCloskey, 2020] McCloskey, A. (2020). Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *arXiv preprint arXiv:2011.12873*.

[McFadden, 1973] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*. edited by P. Zarembka, New York: Wiley.

[McFadden, 1974] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Fontiers in Econometrics*, pages 105–142. Academic press, New York.

[Meinshausen et al., 2009] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al., 2013b] Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013b). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

[Nibbering and Hastie, 2022] Nibbering, D. and Hastie, T. J. (2022). Multiclass-penalized logistic regression. *Computational Statistics & Data Analysis*, 169:107414.

[Palmgren, 1981] Palmgren, J. (1981). The fisher information matrix for log linear models arguing conditionally on observed explanatory variable. *Biometrika*, 68(2):563–566.

[Partalas et al., 2015] Partalas, I., Kosmopoulos, A., Baskiotis, N., Artieres, T., Paliouras, G., Gaussier, E., Androutsopoulos, I., Amini, M.-R., and Galinari, P. (2015). Lshtc: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*.

[Pastorello et al., 2003] Pastorello, S., Patilea, V., and Renault, E. (2003). Iterative and recursive estimation in structural nonadaptive models. *Journal of Business & Economic Statistics*, 21(4):449–509.

[Pellegrini and Fotheringham, 2002] Pellegrini, P. A. and Fotheringham, A. S. (2002). Modelling spatial choice: a review and synthesis in a migration context. *Progress in Human Geography*, 26(4):487–510.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

[Raman et al., 2016] Raman, P., Srinivasan, S., Matsushima, S., Zhang, X., Yun, H., and Vishwanathan, S. (2016). Ds-mlr: exploiting double separability for scaling up distributed multinomial logistic regression. *arXiv preprint arXiv:1604.04706*.

[Reardon and Firebaugh, 2002] Reardon, S. F. and Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological methodology*, 32(1):33–67.

[Recht et al., 2011] Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, 24.

[Rosenbaum, 1984a] Rosenbaum, P. R. (1984a). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666.

[Rosenbaum, 1984b] Rosenbaum, P. R. (1984b). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., and Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

[Shakespeare, 1946] Shakespeare, W. (1946). *Hamlet*. F.S. Crofts & Co., Inc., NY. Act I, Scene 3, Lines 70-72, are apropos.

[Simon et al., 2013] Simon, N., Friedman, J., and Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*.

[Song et al., 2005] Song, P. X.-K., Fan, Y., and Kalbfleisch, J. D. (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100(472):1145–1158.

[Spivak, M.D., Ph.D., 1985] Spivak, M.D., Ph.D. (1985). *PCTEX Manual*. Personal TEX, Inc., CA.

[Spivak, M.D., Ph.D., 1986] Spivak, M.D., Ph.D. (1986). *The Joy of TEX*. American Mathematical Society, RI.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[Taddy, 2013] Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

[Taddy, 2015a] Taddy, M. (2015a). Distributed multinomial regression.

[Taddy, 2015b] Taddy, M. (2015b). Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

[Taddy et al., 2015] Taddy, M., Gardner, M., Chen, L., and Draper, D. (2015). A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation.

[Tang et al., 2010] Tang, D., Agarwal, A., O'Brien, D., and Meyer, M. (2010). Overlapping experiment infrastructure: more, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 17–26, New York, NY, USA. Association for Computing Machinery.

[Tel, 2000] Tel, G. (2000). *Introduction to Distributed Algorithms*. Cambridge university press.

[Tibshirani et al., 2016] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

[Train, 2009] Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

[Vafa et al.,] Vafa, K., Palikot, E., Du, T., Kanodia, A., Athey, S., and Blei, D. Career: Transfer learning for economic prediction of labor data.

[Vafa et al., 2023] Vafa, K., Palikot, E., Du, T., Kanodia, A., Athey, S., and Blei, D. (2023). Career: A foundation model for labor sequence data. *Transactions on Machine Learning Research*.

[van der Vaart, 2000] van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

[Van Steen and Tanenbaum, 2017] Van Steen, M. and Tanenbaum, A. S. (2017). *Distributed Systems*. Maarten van Steen Leiden, The Netherlands.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vincent and Hansen, 2014] Vincent, M. and Hansen, N. R. (2014). Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786.

[Wasserman and Roeder, 2009] Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

[Wilcox, 2012] Wilcox, R. (2012). Chapter 10 - robust regression. In Wilcox, R., editor, *Introduction to Robust Estimation and Hypothesis Testing (Third Edition)*, Statistical Modeling and Decision Science, pages 471–532. Academic Press, Boston, third edition edition.

[Wu, 2018] Wu, A. H. (2018). Gendered language on the economics job market rumors forum. In *AEA Papers and Proceedings*, volume 108, pages 175–179. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

[Yun et al., 2020] Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. (2020). Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*.

[Zhou and Srikumar, 2021] Zhou, Y. and Srikumar, V. (2021). A closer look at how fine-tuning changes bert. *arXiv preprint arXiv:2106.14282*.

Appendix A

ITERATIVE DISTRIBUTED MULTINOMIAL REGRESSION

A.1 Some Notations and Equalities

In this appendix, we list some mathematical expressions and equalities used in the paper.

1. $l_{C|V,M}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i - M_i \log\left(\sum_{k=1}^{d} e^{\eta_{ik}}\right)\right]$
2. $L_{C|V,M}(\boldsymbol{\theta}) \equiv p\lim_{n\to\infty} \frac{1}{n} l_{C|V,M}(\boldsymbol{\theta})$
3. $Q_n^*(\boldsymbol{\theta}) \equiv -l_{C|V,M}(\boldsymbol{\theta})$
4. $\widetilde{\boldsymbol{\theta}} \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*(\boldsymbol{\theta})$
5. $l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu}) \equiv \sum_{i=1}^{n}\left[\boldsymbol{C}_i'(\boldsymbol{\eta}_i + \mu_i \mathbf{1}_d) - M_i \log\left(\sum_{k=1}^{d} e^{\eta_{ik}+\mu_i}\right)\right] = l_{C|V,M}(\boldsymbol{\theta})$
6. $f(\boldsymbol{\theta},\boldsymbol{\mu}) \equiv \sum_{i=1}^{n}\left[M_i \log\left(\sum_{k=1}^{d} e^{\eta_{ik}}\right) + M_i\mu_i - e^{\mu_i}\sum_{k=1}^{d} e^{\eta_{ik}}\right]$
7. $l(\boldsymbol{\theta},\boldsymbol{\mu}) \equiv l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu}) + f(\boldsymbol{\theta},\boldsymbol{\mu}) = \sum_{i=1}^{n}\sum_{k=1}^{d}\left(C_{ik}(\eta_{ik}+\mu_i) - e^{(\eta_{ik}+\mu_i)}\right)$
8. $Q_n(\boldsymbol{\theta},\boldsymbol{\mu}) \equiv -l(\boldsymbol{\theta},\boldsymbol{\mu})$
9. $\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\mu}}\right) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\mu}\in\mathbb{R}^n} Q_n(\boldsymbol{\theta},\boldsymbol{\mu})$,
10. $Q_{kn}(\boldsymbol{\theta}_k,\boldsymbol{\mu}) \equiv \sum_{i=1}^{n}\left(e^{\left(\boldsymbol{V}_i'\boldsymbol{\theta}_k+\mu_i\right)} - C_{ik}(\boldsymbol{V}_i'\boldsymbol{\theta}_k+\mu_i)\right)$
11. $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta}) \equiv \left(\log\left(\frac{M_1}{\sum_{k=1}^{d} e^{\eta_{1k}}}\right),\ldots,\log\left(\frac{M_n}{\sum_{k=1}^{d} e^{\eta_{nk}}}\right)\right)'$
12. $\widehat{\boldsymbol{\theta}}^{(s)} \equiv \left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right),\ldots,\arg\min_{\boldsymbol{\theta}_d\in\Theta_d} Q_{dn}\left(\boldsymbol{\theta}_d, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right)\right]'$

13. $l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right) \equiv \sum_{i=1}^{n}\left[C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k - \left(C_{ik}+C_{id}\right)\log\left(e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}+1\right)\right]$

14. $\check{\boldsymbol{\theta}}_k \equiv \arg\min_{\boldsymbol{\theta}_k\in\Theta_k} -l_{C_k,C_d|V,N_k}\left(\boldsymbol{\theta}_k\right)$

15. $\widehat{\boldsymbol{\theta}}_T \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta},\widehat{\boldsymbol{\mu}}_T\right)$ with $\widehat{\boldsymbol{\mu}}_T = \left(\log\left(M_1\right),\ldots,\log\left(M_n\right)\right)'$

16. $\widehat{\boldsymbol{\theta}}_P \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta},\mathbf{0}\right)$

17. $Q_n^{\dagger}\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \equiv Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\vartheta}\right)\right) = \sum_{i=1}^{n}\sum_{k=1}^{d}\left(\frac{M_i e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}_i'\boldsymbol{\vartheta}_k}} - C_{ik}\boldsymbol{V}_i'\boldsymbol{\theta}_k - C_{ik}\log\left(\frac{M_i}{\sum_{k=1}^{d} e^{V_i'\boldsymbol{\vartheta}_k}}\right)\right)$

18. $Q^{\dagger}\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \equiv \sum_{k=1}^{d}\mathbb{E}\left[\left(\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k - C_k\log\left(\frac{M}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\right)\right)\right]$

19. $Q_k^{\dagger}\left(\boldsymbol{\theta}_k,\boldsymbol{\vartheta}\right) \equiv \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k - C_k\log\left(\frac{M}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\right)\right]$

20. $\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}\right) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\dagger}\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)$

21. $Q^{*}\left(\boldsymbol{\theta}^{*}\right) \equiv p\lim_{n\to\infty}\frac{1}{n}Q_n^{*}\left(\boldsymbol{\theta}^{*}\right)$

22. $\mathcal{I}\left(\boldsymbol{\theta}^{*}\right) \equiv \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q^{*}\left(\boldsymbol{\theta}^{*}\right)$

23. $\frac{\partial^2 Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{bmatrix} \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_d}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] \end{bmatrix}$

24. $\frac{\partial^2 Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'} = \begin{bmatrix} \mathbb{E}\left[-\frac{Me^{2\boldsymbol{V}'\boldsymbol{\theta}_1}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbb{E}\left[-\frac{Me^{\boldsymbol{V}'(\boldsymbol{\theta}_1+\boldsymbol{\theta}_d)}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[-\frac{Me^{\boldsymbol{V}'(\boldsymbol{\theta}_1+\boldsymbol{\theta}_d)}}{\left(\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}\right)^2}\boldsymbol{V}\boldsymbol{V}'\right] & \cdots & \mathbb{E}\left[\frac{Me^{2\boldsymbol{V}'\boldsymbol{\theta}_d}}{\sum_{k=1}^{d} e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}}\boldsymbol{V}\boldsymbol{V}'\right] \end{bmatrix}$

A.2 Proofs

Proof of Lemma 1.2.1: It holds that

$$\begin{aligned} l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu}) &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i+\mu_i\boldsymbol{C}_i'\mathbf{1}_n-M_i\log\left(e^{\mu_i}\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\ &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i+\mu_i\sum_{k=1}^{d}C_{ik}-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)-M_i\mu_i\right] \\ &= \sum_{i=1}^{n}\left[\boldsymbol{C}_i'\boldsymbol{\eta}_i-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\ &= l_{C|V,M}(\boldsymbol{\theta}), \end{aligned}$$

where the second to last equality holds because $\mu_i\sum_{k=1}^{d}C_{ik}=M_i\mu_i$. Therefore, adding μ_i does not change the likelihood. □

Proof of Lemma 1.2.2: Notice that $Q_n(\boldsymbol{\theta},\boldsymbol{\mu})$ is differentiable w.r.t. $\boldsymbol{\mu}$ for any given $\boldsymbol{\theta}$. By letting $\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}=\mathbf{0}$, we can obtain function $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$ such that $\left.\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})}=\mathbf{0}$ for every $\boldsymbol{\theta}$. By definition,

$$Q_n(\boldsymbol{\theta},\boldsymbol{\mu})=-\left[l_{C|V,M}(\boldsymbol{\theta},\boldsymbol{\mu})+f(\boldsymbol{\theta},\boldsymbol{\mu})\right]=-\left[l_{C|V,M}(\boldsymbol{\theta})+f(\boldsymbol{\theta},\boldsymbol{\mu})\right],$$

which implies that $\frac{\partial Q_n(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}=\frac{\partial f(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\boldsymbol{\mu}}$. Since $\frac{\partial f(\boldsymbol{\theta},\boldsymbol{\mu})}{\partial\mu_i}=M_i-e^{\mu_i}\sum_{k=1}^{d}e^{\eta_{ik}}$, we obtain the expression of $\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})$ as in (1.3.3). Plugging it into (1.2.5), we have that

$$\begin{aligned} & f(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n(\boldsymbol{\theta})) \\ =\ & -\sum_{i=1}^{n}\left[M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)+M_i\log\left(\frac{m_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)-e^{\log\left(\frac{M_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)}\sum_{k=1}^{d}e^{\eta_{ik}}\right] \\ =\ & -\sum_{i=1}^{n}\left[M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)+M_i\log M_i-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)-\left(\frac{M_i}{\sum_{k=1}^{d}e^{\eta_{ik}}}\right)\sum_{k=1}^{d}e^{\eta_{ik}}\right] \\ =\ & -\sum_{i=1}^{n}\left[M_i\log M_i-M_i\right], \end{aligned} \tag{A.2.1}$$

which does not depend on $\boldsymbol{\theta}$. As a result, it holds that

$$\arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\theta}\right)\right) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta}\left[Q_n^*\left(\boldsymbol{\theta}\right) - f\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\boldsymbol{\theta}\right)\right)\right] = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*\left(\boldsymbol{\theta}\right).$$

Because $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}\left(\boldsymbol{\theta}\right)\right)$, we have that

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_n^*\left(\boldsymbol{\theta}\right) = \widetilde{\boldsymbol{\theta}}.$$

The claimed lemma then follows. □

Proof of Lemma 1.3.2: Under the assumption that $\Pr\left(M_i \mid \boldsymbol{V}_i\right) = \text{Po}\left(\sum_{k=1}^{d} e^{\eta_{ik}^*}\right)$, we can obtain that

$$\begin{aligned}\Pr\left(\boldsymbol{C}_i \mid \boldsymbol{V}_i\right) &= \Pr\left(\boldsymbol{C}_i \mid \boldsymbol{V}_i, M_i\right) \text{Po}\left(\sum_{k=1}^{d} e^{\eta_{ik}^*}\right) \\ &= \prod_{k=1}^{d} \text{Po}\left(e^{\eta_{ik}^*}\right) = \prod_{k=1}^{d} \frac{e^{\eta_{ik}^* C_{ik}} e^{-e^{\eta_{ik}^*}}}{C_{ik}!} = \frac{\prod_{k=1}^{d} e^{\eta_{ik}^* C_{ik}}}{C_{i1}! \cdots C_{id}!} e^{-\sum_{k=1}^{d} e^{\eta_{ik}^*}}.\end{aligned}$$

The log likelihood function is written as

$$\begin{aligned}\log\left[\prod_{i=1}^{n} \frac{\prod_{k=1}^{d} e^{\eta_{ik} C_{ik}}}{C_{i1}! \cdots C_{id}!} e^{-\sum_{k=1}^{d} e^{\eta_{ik}}}\right] &\sim \sum_{i=1}^{n}\left[\sum_{k=1}^{d} C_{ik}\eta_{ik} - \sum_{k=1}^{d} e^{\eta_{ik}}\right] = \sum_{i=1}^{n}\sum_{k=1}^{d}\left(C_{ik}\eta_{ik} - e^{\eta_{ik}}\right) \\ &= l\left(\boldsymbol{\theta}, \mathbf{0}\right).\end{aligned}$$

Therefore, $\widehat{\boldsymbol{\theta}}_P$ maximizes the true log likelihood function based upon $\Pr\left(\boldsymbol{C}_i \mid \boldsymbol{V}_i\right)$. The lemma follows. □

Proof of Lemma 1.3.1: By Equation (1.2.1), it holds that

$$\begin{aligned}&\Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, M_i\right)\\ =&\frac{M_i!}{C_{ik}!C_{id}!\left(M_i - C_{ik} - C_{id}\right)!}\left(\frac{e^{\eta_{ik}^*}}{\Lambda_i^*}\right)^{C_{ik}}\left(\frac{e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{C_{id}}\left(\frac{\Lambda_i^* - e^{\eta_{ik}^*} - e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{M_i - C_{ik} - C_{id}}. \end{aligned} \tag{A.2.2}$$

Because $N_{ik} = C_{ik} + C_{id}$, we have that

$$\begin{aligned}\Pr\left(C_{ik}, C_{id} \mid \boldsymbol{V}_i, M_i\right) &= \Pr\left(C_{ik}, C_{id}, N_{ik} \mid \boldsymbol{V}_i, M_i\right)\\ &= \Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i, M_i\right)\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right). \end{aligned} \tag{A.2.3}$$

Compute $\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right)$ as

$$\Pr\left(N_{ik} \mid \boldsymbol{V}_i, M_i\right) = \frac{M_i!}{N_{ik}!\left(M_i - N_{ik}\right)!}\left(\frac{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{N_{ik}}\left(\frac{\Lambda_i^* - e^{\eta_{ik}^*} - e^{\eta_{id}^*}}{\Lambda_i^*}\right)^{M_i - N_{ik}}.$$

Together with Equations (A.2.2) and (A.2.3), we obtain that

$$\begin{aligned}\Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i, M_i\right) &= \frac{N_{ik}!}{C_{ik}!C_{id}!}\left(\frac{e^{\eta_{ik}^*}}{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}\right)^{C_{ik}}\left(\frac{e^{\eta_{id}^*}}{e^{\eta_{ik}^*} + e^{\eta_{id}^*}}\right)^{C_{id}}\\ &= \Pr\left(C_{ik}, C_{id} \mid N_{ik}, \boldsymbol{V}_i\right).\end{aligned}$$

By replacing $e^{\eta_{id}^*}$ with 1 because $\boldsymbol{\theta}_d^* = \mathbf{0}$, we obtain the claimed result. □

Lemma A.2.1. *Under Assumption 1.4.2, $-L_{C|V,M}\left(\boldsymbol{\theta}\right)$ is convex in $\boldsymbol{\theta}$.*

Proof of Lemma A.2.1: By the definitions of $l_{C|V,M}(\boldsymbol{\theta})$ and $L_{C|V,M}(\boldsymbol{\theta})$ and Assumption 1.4.2, we have that

$$\begin{aligned}
-\frac{1}{n}l_{C|V,M}(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{i=1}^{n}\left[C_{i1}\eta_{i1}+\cdots+C_{id}\eta_{id}-M_i\log\left(\sum_{k=1}^{d}e^{\eta_{ik}}\right)\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[C_{i1}\boldsymbol{V}_i'\boldsymbol{\theta}_1+\cdots+C_{id}\boldsymbol{V}_i'\boldsymbol{\theta}_d-M_i\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}_i'\boldsymbol{\theta}_k}\right)\right] \\
&\xrightarrow{p} -\left(\mathbb{E}\left[C_1\boldsymbol{V}'\boldsymbol{\theta}_1\right]+\cdots+\mathbb{E}\left[C_d\boldsymbol{V}'\boldsymbol{\theta}_d\right]+\mathbb{E}\left[M\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)\right]\right) \\
&\equiv -L_{C|V,M}(\boldsymbol{\theta}) \\
&= -\sum_{k=1}^{d}\mathbb{E}\left[C_k\boldsymbol{V}'\boldsymbol{\theta}_k\right]+\mathbb{E}\left[M\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)\right].
\end{aligned}$$

The first term is convex in $\boldsymbol{\theta}$ because it only involves linear functions. It has been shown in Section 3.1.5 in [Boyd et al., 2004] that $\log\left(\sum_{k=1}^{d}e^{\boldsymbol{V}'\boldsymbol{\theta}_k}\right)$ is convex in $\boldsymbol{V}'\boldsymbol{\theta}_1,\ldots,\boldsymbol{V}'\boldsymbol{\theta}_d$. Because $\boldsymbol{V}'\boldsymbol{\theta}_k$ is a linear function of θ_k and sums of convex functions are convex, the second term is convex in $\boldsymbol{\theta}$. □

Lemma A.2.2. *$\boldsymbol{\theta}^*$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\boldsymbol{\theta})$.*

Proof of Lemma A.2.2: By definition, $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})$. Thus, if we can show that $\frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \mathbf{0}$ holds only at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, then $\boldsymbol{\theta}^*$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\cdot)$. Taking the first order derivative of $-L_{C|V,M}(\cdot)$, we obtain that for any $\check{\boldsymbol{\theta}}\in\Theta$,

$$\begin{aligned}
-\frac{d}{d\boldsymbol{\theta}}L_{C|V,M}(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} &= \left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\check{\boldsymbol{\theta}}_1}\boldsymbol{V}'}{\sum_{k=1}^{d}e^{\boldsymbol{V}'\check{\boldsymbol{\theta}}_k}}-C_1\boldsymbol{V}'\right],\ldots,\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\check{\boldsymbol{\theta}}_{d-1}}\boldsymbol{V}'}{\sum_{k=1}^{d}e^{\boldsymbol{V}'\check{\boldsymbol{\theta}}_k}}-C_{d-1}\boldsymbol{V}'\right]\right]' \\
&= \frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=(\check{\boldsymbol{\theta}},\check{\boldsymbol{\theta}})}.
\end{aligned}$$

Therefore, proving that $\frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}} \mid_{\boldsymbol{\vartheta}=\boldsymbol{\theta}} = \mathbf{0}$ holds only at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ is equivalent to showing that $\frac{d}{d\boldsymbol{\theta}}L_{C|V,M}(\boldsymbol{\theta}) = \mathbf{0}$ only at $\boldsymbol{\theta}^*$. By Lemma A.2.1, $-L_{C|V,M}(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. Therefore, $-L_{C|V,M}(\boldsymbol{\theta})$ only has minimums in the interior of Θ. In addition, for a convex function over

a convex set, any local minimum is also a global minimum. Along with the identification assumption that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} -L_{C|V,M}(\boldsymbol{\theta})$, we obtain that $\frac{\partial}{\partial\boldsymbol{\theta}} L_{C|V,M}(\boldsymbol{\theta}) = \mathbf{0}$ holds only at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. □

Proof of Lemma 1.4.1: Part (i) follows by applying Theorem 2 in [Jennrich, 1969] on the uniform law of large numbers with conditions satisfied by Assumption 1.4.2.

Since $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is continuous in $\boldsymbol{\theta}$ and Θ is compact by Assumption 1.4.2 (i), $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ achieves its minimum in Θ for any $\boldsymbol{\vartheta}$. Given any $0 < \lambda < 1$ and $\boldsymbol{\theta}^1 \neq \boldsymbol{\theta}^2$, if $v_j \neq 0$ for $j = 1, \ldots, p$, then we have

$$e^{\boldsymbol{v}'\left[\lambda\boldsymbol{\theta}^1 + (1-\lambda)\boldsymbol{\theta}^2\right]} < \lambda e^{\boldsymbol{v}'\boldsymbol{\theta}^1} + (1-\lambda)\, e^{\boldsymbol{v}'\boldsymbol{\theta}^2}.$$

In consequence, $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta}$. Because Θ is a convex set, we have that $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ have a unique minimizer for any $\boldsymbol{\vartheta}$. Part (ii) holds.

For part (iii), because $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta} \in \Theta$ and Θ is convex, the (opposite) maximum theorem implies the continuity.

To prove part (iv), we take the first order partial derivative of $Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ with respect to $\boldsymbol{\theta}$ and obtain that

$$\frac{\partial Q^\dagger(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \left[\mathbb{E}\left[\frac{M e^{\boldsymbol{V}'\boldsymbol{\theta}_1}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_1\boldsymbol{V}'\right], \ldots, \mathbb{E}\left[\frac{M e^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\vartheta}_k}} - C_{d-1}\boldsymbol{V}'\right]\right]'.$$

Since $\mathbb{E}\left[C_k \mid \boldsymbol{V}, M\right] = \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}$, we have that

$$\begin{aligned}
&\frac{\partial Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right)}{\partial \boldsymbol{\theta}} \mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)} \\
&= \left[\mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - C_1\boldsymbol{V}' \mid \boldsymbol{V}, M\right]\right], \ldots, \mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - C_d\boldsymbol{V}' \mid \boldsymbol{V}, M\right]\right]\right]' \\
&= \left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\right], \ldots, \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}} - \frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_{d-1}^*}\boldsymbol{V}'}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\right]\right]' \\
&= \mathbf{0}. \qquad \text{(A.2.4)}
\end{aligned}$$

Because $Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\vartheta}\right)$ is strictly convex in $\boldsymbol{\theta}$ for any $\boldsymbol{\vartheta}$, $Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right)$ is certainly strictly convex in $\boldsymbol{\theta}$. Combining with (A.2.4), we obtain that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} Q^\dagger\left(\boldsymbol{\theta}, \boldsymbol{\theta}^*\right)$, which implies that $\overline{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^*\right) = \boldsymbol{\theta}^*$.

Part (v) is proved by Lemma A.2.2. □

Proof of Theorem 1.4.2: For part (i), we show that if $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$ as $n \to \infty$, then $\widehat{\boldsymbol{\theta}}^{(1)} \xrightarrow{p} \boldsymbol{\theta}^*$ as well. By (1.3.4), $\widehat{\boldsymbol{\theta}}_1^{(1)}$ satisfies that

$$\widehat{\boldsymbol{\theta}}_1^{(1)} = \arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right).$$

The first order condition provides that

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\widehat{\boldsymbol{\theta}}_1^{(1)}} = \mathbf{0}.$$

The mean value theorem implies that

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}_1} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^*} + \frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'} Q_{1n}\left(\boldsymbol{\theta}_1, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(0)}\right)\right) \mid_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_1^\star} \left(\widehat{\boldsymbol{\theta}}_1^{(1)} - \boldsymbol{\theta}_1^*\right) \\
&\equiv A_n + B_b\left(\widehat{\boldsymbol{\theta}}_1^{(1)} - \boldsymbol{\theta}_1^*\right),
\end{aligned}$$

where $\boldsymbol{\theta}_1^\star$ lies between $\widehat{\boldsymbol{\theta}}_1^{(1)}$ and $\boldsymbol{\theta}_1^*$. Since $\widehat{\boldsymbol{\theta}}^{(0)} \xrightarrow{p} \boldsymbol{\theta}^*$, we have that

$$A_n \xrightarrow{p} \mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\boldsymbol{V} - C_k\boldsymbol{V}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Me^{\boldsymbol{V}'\boldsymbol{\theta}_1^*}}{\sum_{k=1}^d e^{\boldsymbol{V}'\boldsymbol{\theta}_k^*}}\boldsymbol{V} - C_k\boldsymbol{V} \mid \boldsymbol{V}, M\right]\right] = \mathbf{0}.$$

By a similar argument, it can be shown that B_n converges in probability to a non-singular matrix for any $\boldsymbol{\theta}_1^\star \in \Theta_1$. Therefore, it must hold that $\widehat{\boldsymbol{\theta}}_1^{(1)} \xrightarrow{p} \boldsymbol{\theta}_1^*$. Hence, $\widehat{\boldsymbol{\theta}}^{(S)} \xrightarrow{p} \boldsymbol{\theta}^*$ as $n \to \infty$ for any S.

We now prove part (ii) of the theorem. Based on Lemma 1.4.1, Assumptions 1, 2a, and 5 in [Pastorello et al., 2003] are satisfied. Therefore, Proposition 1 in [Pastorello et al., 2003] holds, which implies that as $n \to \infty$,

$$\sup_{\boldsymbol{\vartheta}\in\Theta} \left\|\overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \xrightarrow{p} 0, \tag{A.2.5}$$

where $\overline{\boldsymbol{\theta}}_n(\cdot)$ is defined in (1.3.4).

Let $\boldsymbol{\vartheta}^{(0)} \equiv p\lim_{n\to\infty} \widehat{\boldsymbol{\theta}}^{(0)}$, where $\widehat{\boldsymbol{\theta}}^{(0)}$ is the initial value of our IDC estimator $\widehat{\boldsymbol{\theta}}^I$. It can be seen that

$$\boldsymbol{\vartheta}^{(0)} = \left[\arg\min_{\boldsymbol{\theta}_1\in\Theta_1} Q_1(\boldsymbol{\theta}_1), \ldots, \arg\min_{\boldsymbol{\theta}_d\in\Theta_d} Q_d(\boldsymbol{\theta}_d)\right]',$$

where $Q_k(\boldsymbol{\theta}_k) \equiv \mathbb{E}\left[e^{\boldsymbol{V}'\boldsymbol{\theta}_k} - C_k\boldsymbol{V}'\boldsymbol{\theta}_k\right]$. Define $\boldsymbol{\vartheta}^{(1)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(0)}\right)$, $\boldsymbol{\vartheta}^{(2)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(1)}\right) \equiv \overline{\boldsymbol{\theta}}^2\left(\boldsymbol{\vartheta}^{(0)}\right)$, ..., $\boldsymbol{\vartheta}^{(s)} \equiv \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(s-1)}\right) \equiv \overline{\boldsymbol{\theta}}^s\left(\boldsymbol{\vartheta}^{(0)}\right)$ for any $s \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ denotes the set of positive integers. Next, we show that $\left(\boldsymbol{\vartheta}^{(s)}\right)$ is a Cauchy sequence. By Assumption 1.4.3, we have that for any

$s_1 > s_2 \geq 1$,

$$\begin{aligned}
\left\|\boldsymbol{\vartheta}^{(s_1)} - \boldsymbol{\vartheta}^{(s_2)}\right\| &= \left\|\overline{\boldsymbol{\theta}}^{s_1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\| \\
&\leq \left[\left\|\overline{\boldsymbol{\theta}}^{s_1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_1-1}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\| + \left\|\overline{\boldsymbol{\theta}}^{s_1-1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_1-2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\|\right. \\
&\quad \left. + \cdots + \left\|\overline{\boldsymbol{\theta}}^{s_2+1}\left(\boldsymbol{\vartheta}^{(0)}\right) - \overline{\boldsymbol{\theta}}^{s_2}\left(\boldsymbol{\vartheta}^{(0)}\right)\right\|\right] \\
&\leq \left[C^{s_1-1} + C^{s_1-2} + \cdots + C^{s_2}\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&= C^{s_2}\left[\sum_{i=0}^{s_1-s_2-1} C^i\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&\leq C^{s_2}\left[\sum_{i=0}^{\infty} C^i\right]\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\| \\
&\leq \frac{C^{s_2}}{1-C}\left\|\boldsymbol{\vartheta}^{(1)} - \boldsymbol{\vartheta}^{(0)}\right\|, \qquad \text{(A.2.6)}
\end{aligned}$$

which implies that $\left(\boldsymbol{\vartheta}^{(s)}\right)$ is Cauchy because $C < 1$. Since $\Theta \subseteq \mathbb{R}^{p\times d}$ is compact by Assumption 1.4.2 and $\mathbb{R}^{p\times d}$ is complete with respect to $\|\cdot\|$, Θ is also complete with respect to $\|\cdot\|$. Therefore, $\boldsymbol{\vartheta}^{(s)}$ converges to a limit $\boldsymbol{\vartheta}^\star$ in Θ as $s \to \infty$. Because

$$\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^\star\right) = \overline{\boldsymbol{\theta}}\left(\lim_{s\to\infty}\boldsymbol{\vartheta}^{(s)}\right) = \lim_{s\to\infty}\overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(s)}\right) = \lim_{s\to\infty}\boldsymbol{\vartheta}^{(s+1)} = \boldsymbol{\vartheta}^\star,$$

it holds that $\boldsymbol{\vartheta}^\star$ is a fixed point of the mapping $\overline{\boldsymbol{\theta}} : \Theta \to \Theta$. By Lemma A.2.2, $\boldsymbol{\theta}^*$ is the unique fixed point of $\overline{\boldsymbol{\theta}}(\cdot)$. Thus, $\boldsymbol{\vartheta}^\star = \boldsymbol{\theta}^*$ and $\lim_{s\to\infty}\boldsymbol{\vartheta}^{(s)} = \boldsymbol{\theta}^*$.

We now show that $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\vartheta}^{(s)} = o_p(1)$ for any $s \in \mathbb{Z}^+$ by induction. By the definition of $\boldsymbol{\vartheta}^{(0)}$, $\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\vartheta}^{(0)} = o_p(1)$. Assuming that $\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\vartheta}^{(t)} = o_p(1)$ for some t, it holds that

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\vartheta}^{(t+1)} &= \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(t)}\right) \\
&= \left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)\right] + \left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\boldsymbol{\vartheta}^{(t)}\right)\right] \\
&= o_p(1),
\end{aligned}$$

where the last equality holds because $\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) = o_p(1)$ by (A.2.5), $\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\vartheta}^{(t)} = o_p(1)$ by assumption and $\overline{\boldsymbol{\theta}}(\cdot)$ is continuous by Lemma 1.4.1 (iii). Therefore, $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\vartheta}^{(s)} = o_p(1)$ for any $s \in \mathbb{Z}^+$.

Hence, we have that if $S \to \infty$ and $n \to \infty$, then

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}^{I} - \boldsymbol{\theta}^*\right\| &= \left\|\widehat{\boldsymbol{\theta}}^{(S)} - \boldsymbol{\vartheta}^{(S)}\right\| + \left\|\boldsymbol{\vartheta}^{(S)} - \boldsymbol{\theta}^*\right\| \\ &\equiv A(n,S) + B(S) \xrightarrow{p} 0,\end{aligned}$$

because $A(n,S) \xrightarrow{p} 0$ as $n \to \infty$ for any given S and $B(S) \to 0$ as $S \to \infty$ by $\lim_{s\to\infty} \boldsymbol{\vartheta}^{(s)} = \boldsymbol{\theta}^*$. The second part of the theorem holds. □

Lemma A.2.3. *Under the conditions in Theorem 1.4.3, it holds that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right) \xrightarrow{p} 0.$$

Proof of Lemma A.2.3: We first show that the result in the lemma holds if the conditions in part (i) of the theorem hold. By (1.3.4), for any $s \in \mathbb{Z}^+$, $\widehat{\boldsymbol{\theta}}^{(s)}$ satisfies that

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(s-1)}\right)\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}} = \frac{\partial}{\partial \boldsymbol{\theta}} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} = \mathbf{0}.$$

Apply Taylor expansion to the left-hand-side of the equality at $\boldsymbol{\theta}^*$. Because $\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^* = o_p(1)$ and $\widehat{\boldsymbol{\theta}}^{(s-1)} - \boldsymbol{\theta}^* = o_p(1)$, we obtain that

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\theta}} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*, \boldsymbol{\vartheta}=\boldsymbol{\theta}^*} + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} \left(\widehat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^*\right) \\ &+ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\vartheta}'} Q_n^{\dagger}(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)}, \boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}} \left(\widehat{\boldsymbol{\theta}}^{(s-1)} - \boldsymbol{\theta}^*\right) = \mathbf{0} \qquad \text{(A.2.7)}\end{aligned}$$

by ignoring higher order terms. Since $\frac{\partial^2 Q^\dagger(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}$ is non-singular and both $\widehat{\boldsymbol{\theta}}^{(s)}$ and $\widehat{\boldsymbol{\theta}}^{(s-1)}$ are consistent estimators of $\boldsymbol{\theta}^*$, we have that

$$\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}$$

exists with high probability when n is large. Define

$$\begin{aligned}
A_n &= \left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}\left(-\frac{\partial}{\partial\boldsymbol{\theta}}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)\\
B_n &= \left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right)^{-1}\left(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(s)},\boldsymbol{\vartheta}=\widehat{\boldsymbol{\theta}}^{(s-1)}}\right).
\end{aligned}$$

By the law of large numbers and the consistency of $\widehat{\boldsymbol{\theta}}^{(s)}$ for any s, we have that

$$\begin{aligned}
A_n &= \left(\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)^{-1}\left(-\frac{\partial}{\partial\boldsymbol{\theta}}Q_n^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)+o_p\left(1\right)\equiv A+o_p\left(1\right)\\
B_n &= \left(\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)^{-1}\left(-\frac{\partial^2 Q^\dagger\left(\boldsymbol{\theta},\boldsymbol{\vartheta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\vartheta}'} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\boldsymbol{\vartheta}=\boldsymbol{\theta}^*}\right)+o_p\left(1\right)\equiv B+o_p\left(1\right).
\end{aligned}$$

By ignoring the smaller order terms, we obtain from Equation (A.2.7) that

$$\widehat{\boldsymbol{\theta}}^{(s)}-\boldsymbol{\theta}^*=A+B\left(\widehat{\boldsymbol{\theta}}^{(s-1)}-\boldsymbol{\theta}^*\right)=\sum_{t=0}^{s-1}B^tA+B^s\left(\widehat{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right),$$

where the second equality follows from iterating the first equality. It then holds that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}\right)=\sqrt{n}B^SA+\sqrt{n}B^S\left(B-I\right)\left(\widehat{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right).$$

By Assumption 1.4.4, $\|B\|<1$, which implies that $\sqrt{n}\left\|B\right\|^S\to 0$ as $S\geq\log\left(n\right)$ and $n\to\infty$. Hence, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}\right)=o_p\left(1\right)$. The claimed lemma follows.

We now prove that under the conditions in part (ii) of the theorem, the result also holds. If we can show that for any $s\in\mathbb{Z}^+$, with probability approaching one there exists a constant

$c < 1$ such that

$$\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \leq c\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|, \tag{A.2.8}$$

then by the same derivation as (A.2.6), we would obtain that

$$\left\|\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right\| \leq \frac{c^S}{1-c}\left\|\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}^{(0)}\right\|.$$

Because $c < 1$ and $S > n^{\delta}$ for some $\delta > 0$, it holds that

$$\sqrt{n}\left\|\widehat{\boldsymbol{\theta}}^{(S+1)} - \widehat{\boldsymbol{\theta}}^{(S)}\right\| \leq \frac{n^{\frac{1}{2}}c^S}{1-c}\left\|\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}^{(0)}\right\| \to 0.$$

Thus, it suffices to prove (A.2.8).

By the implicit function theorem, $\overline{\boldsymbol{\theta}}(\cdot)$ is continuously differentiable in Θ. Together with Assumption 1.4.3, we have that there exists $\epsilon > 0$ such that for any $\boldsymbol{\vartheta} \in \Theta^d$ and $\check{\boldsymbol{\vartheta}} \in \mathcal{B}^{\epsilon}\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right) \equiv \left\{\boldsymbol{\theta} \in \Theta^d : \left\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \leq \epsilon\right\}$, we have

$$\left\|\overline{\boldsymbol{\theta}}(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\right\| \leq C_{\epsilon}\left\|\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\right\|,$$

for some $C_{\epsilon} \leq \overline{C} < 1$. By (A.2.5), $\Pr\left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right) \in \mathcal{B}^{\epsilon}\left(\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right)\right] \to 1$ as $n \to \infty$ for any s and ϵ. Therefore, with probability approaching one, it holds that

$$\begin{aligned}\left\|\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| &= \left\|\overline{\boldsymbol{\theta}}\left(\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \\ &\leq C_{\epsilon}\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right) - \widehat{\boldsymbol{\theta}}^{(s)}\right\| = C_{\epsilon}\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|.\end{aligned} \tag{A.2.9}$$

For any $\boldsymbol{\vartheta} \in \Theta^d$, define $\Omega(\boldsymbol{\vartheta}) \equiv \frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$, where $\frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$ is the Jacobian matrix of dimension $dp \times dp$. By the implicit function theorem, we have

$$\frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) = \left[\frac{\partial}{\partial \boldsymbol{\theta}}g(\boldsymbol{\theta}, \boldsymbol{\vartheta})\mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\right)}\right]^{-1}\frac{\partial}{\partial \boldsymbol{\vartheta}}g(\boldsymbol{\theta}, \boldsymbol{\vartheta})\mid_{(\boldsymbol{\theta},\boldsymbol{\vartheta})=\left(\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\right)},$$

where $g(\boldsymbol{\theta},\boldsymbol{\vartheta}) \equiv \frac{\partial Q^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}$ is the function that defines $\overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta})$. Similarly, we define $\Omega_n(\boldsymbol{\vartheta}) \equiv \frac{\partial}{\partial \boldsymbol{\vartheta}}\overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta})$ and $g_n(\boldsymbol{\theta},\boldsymbol{\vartheta}) \equiv \frac{\partial Q_n^{\dagger}(\boldsymbol{\theta},\boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}}$. By Assumption 1.4.2 (ii) and (iii) and the uniform law of large numbers, we have

$$\sup_{\boldsymbol{\vartheta}\in\Theta} \|\Omega_n(\boldsymbol{\vartheta}) - \Omega(\boldsymbol{\vartheta})\| \xrightarrow{p} 0. \tag{A.2.10}$$

By Assumption 1.4.2 (i), Θ is convex. Applying a multivariate Taylor expansion ([Dieudonné, 2011], p. 190), we can write $\overline{\boldsymbol{\theta}}(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}(\boldsymbol{\vartheta}) = \Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ for any $\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}} \in \Theta$, where $\Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) \equiv \int_0^1 \Omega(\boldsymbol{\vartheta} + \xi(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}))\, d\xi$. Similarly, we have $\overline{\boldsymbol{\theta}}_n(\check{\boldsymbol{\vartheta}}) - \overline{\boldsymbol{\theta}}_n(\boldsymbol{\vartheta}) = \Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta})$ for any $\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}} \in \Theta^d$, where $\Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) \equiv \int_0^1 \Omega_n(\boldsymbol{\vartheta} + \xi(\check{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}))\, d\xi$. It holds that

$$\begin{aligned}
\left\|\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| &\leq \left\|\left[\overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}_n\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right] - \left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right]\right\| \\
&\quad + \left\|\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right\| \\
&\leq \left\|\left[\Lambda_n(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}}) - \Lambda(\boldsymbol{\vartheta},\check{\boldsymbol{\vartheta}})\right]\left[\overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s+1)}\right) - \overline{\boldsymbol{\theta}}\left(\widehat{\boldsymbol{\theta}}^{(s)}\right)\right]\right\| \\
&\quad + C_\epsilon \left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|,
\end{aligned}$$

where the first inequality follows from the triangular inequality and the second inequality holds by (A.2.9). Because the first term on the right hand side has the order $o_p\left(\left\|\widehat{\boldsymbol{\theta}}^{(s+1)} - \widehat{\boldsymbol{\theta}}^{(s)}\right\|\right)$ because of (A.2.10), we have shown that (A.2.8) holds with $c = C_\epsilon < 1$. The lemma follows. □

Proof of Theorem 1.4.3: We aim to show that $\widehat{\boldsymbol{\theta}}^{(S)}$ has the same influence function as $\widetilde{\boldsymbol{\theta}}$. By the definition of the IDC estimator in Section **??**, we have that

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q_n\left(\boldsymbol{\theta}, \overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S)}} = \mathbf{0}.$$

Applying the Taylor expansion to the function on the left-hand-side of the above equation round $\widehat{\boldsymbol{\theta}}^{(S-1)}$, we can obtain that

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}\\&+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}}\left(\widehat{\boldsymbol{\theta}}^{(S)}-\widehat{\boldsymbol{\theta}}^{(S-1)}\right)=\mathbf{0},\end{aligned}\tag{A.2.11}$$

where $\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}$ lies between $\widehat{\boldsymbol{\theta}}^{(S)}$ and $\widehat{\boldsymbol{\theta}}^{(S-1)}$. The definition of $Q_n(\boldsymbol{\theta},\boldsymbol{\mu})$ implies that

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}\\&=-\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)-\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}.\end{aligned}$$

Applying the expression of $f(\boldsymbol{\theta},\boldsymbol{\mu})$, it can be shown that

$$\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(S-1)}}=\mathbf{0}.$$

Therefore, (A.2.11) can be rewritten as

$$-\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q_n\left(\boldsymbol{\theta},\overline{\boldsymbol{\mu}}_n\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)\right)|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\dagger(S-1)}}\left(\widehat{\boldsymbol{\theta}}^{(S)}-\widehat{\boldsymbol{\theta}}^{(S-1)}\right)=\mathbf{0}.$$

By Lemma A.2.3, we have $\widehat{\boldsymbol{\theta}}^{(S+1)}-\widehat{\boldsymbol{\theta}}^{(S)}=o_p\left(n^{-1/2}\right)$. This implies that

$$\begin{aligned}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\widehat{\boldsymbol{\theta}}^{(S-1)}\right)&=o_p\left(n^{-1/2}\right)\\&=\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^*\right)+\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^\dagger\right)\left(\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^*\right),\end{aligned}\tag{A.2.12}$$

where the second equality follows from the Taylor expansion and $\boldsymbol{\theta}^\dagger$ lies between $\widehat{\boldsymbol{\theta}}^{(S-1)}$ and $\boldsymbol{\theta}^*$.

By Theorem 1.4.2, we have that $\widehat{\boldsymbol{\theta}}^{(S-1)}\xrightarrow{p}\boldsymbol{\theta}^*$ as $n\to\infty$. Therefore, by Assumption 1.4.2

and Taylor's theorem, we can obtain that

$$\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{\dagger}\right)=\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(1\right).$$

Since matrix inversion is continuous (at non-singular matrices), it follows that the inverse of $\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)$ exists with high probability and

$$\left[-\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)\right]^{-1}\xrightarrow{p}\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right),$$

where $\mathcal{I}\left(\boldsymbol{\theta}^{*}\right)$ is the Fisher information matrix defined in Section 1.4.2. Using this result to (A.2.12), we obtain that

$$\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^{*}=\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right)\frac{1}{n}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(n^{-1/2}\right).$$

Since $\widehat{\boldsymbol{\theta}}^{(S)}-\boldsymbol{\theta}^{*}=\widehat{\boldsymbol{\theta}}^{(S-1)}-\boldsymbol{\theta}^{*}+o_p\left(n^{-1/2}\right)$ by Lemma A.2.3, it holds that

$$\widehat{\boldsymbol{\theta}}^{(S)}-\boldsymbol{\theta}^{*}=\mathcal{I}^{-1}\left(\boldsymbol{\theta}^{*}\right)\frac{1}{n}\frac{d}{d\boldsymbol{\theta}}l_{C|V,M}\left(\boldsymbol{\theta}^{*}\right)+o_p\left(n^{-1/2}\right).$$

It can be seen that $\widehat{\boldsymbol{\theta}}^{(S)}$ and the maximum likelihood estimator $\widetilde{\boldsymbol{\theta}}$ have the same influence function. Hence, under the assumptions in either part (i) or (ii) of the theorem, we have $\widehat{\boldsymbol{\theta}}^{(S)}-\widetilde{\boldsymbol{\theta}}=o_p\left(n^{-1/2}\right)$. □

Proof of Corollary 1.4.4: The result directly follows from Theorem 1.4.3 and the standard result on the asymptotic distribution of the maximum likelihood estimator. □

Proof of Theorem 1.4.5: The proof of the theorem follows from the discussion in Section 1.4.2. □

Appendix B

EXAMINING POWER IN CLUSTERED RANDOMIZED PRICING EXPERIMENTS IN E-COMMERCE

B.1 Intra-cluster Correlation Coefficient

We consider two observations Y_{ij} and Y_{kj} from the same cluster j. These observations are modeled as follows:

$$Y_{ij} = \alpha + u_j + \epsilon_{ij}$$
$$Y_{kj} = \alpha + u_j + \epsilon_{kj}$$

We aim to establish that the intra-cluster correlation coefficient (ICC) is equal to the correlation between these two observations. The correlation is defined by:

$$Corr(Y_{ij}, Y_{kj}) = \frac{Cov(Y_{ij}, Y_{kj})}{\sqrt{Var(Y_{ij}) \cdot Var(Y_{kj})}}$$

Considering that Y_{ij} and Y_{kj} share the same cluster, they share the random effect u_j. We calculate the covariance between them as:

$$
\begin{aligned}
Cov(Y_{ij}, Y_{kj}) &= Cov(\alpha + u_j + \epsilon_{ij}, \alpha + u_j + \epsilon_{kj}) \\
&= Cov(u_j, u_j) + Cov(u_j, \epsilon_{kj}) + Cov(\epsilon_{ij}, u_j) + Cov(\epsilon_{ij}, \epsilon_{kj}) \\
&= Var(u_j) + 0 + 0 + 0 \\
&= Var(u_j) \\
&= \sigma_u^2
\end{aligned}
$$

The last three covariance terms vanish because u_j, ϵ_{ij}, and ϵ_{kj} are independent. The variance for both Y_{ij} and Y_{kj} is the total variance, which is the sum of the individual variance and the variance due to the cluster effect:

$$Var(Y_{ij}) = Var(Y_{kj}) = \sigma_y^2 = \sigma_u^2 + \sigma_\epsilon^2$$

Substituting the values of covariance and variance into the correlation formula gives us:

$$Corr(Y_{ij}, Y_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{\sigma_u^2}{\sigma_y^2} = \rho = ICC$$

This concludes the proof that the intra-cluster correlation coefficient (ICC), ρ, is equal to the correlation between two different observations from the same cluster.

We also demonstrate that the intra-class correlation (ICC) increases the variance of treatment effect estimates in cluster-randomized experiments (CRE), we begin with the basic variance formula for the treatment effect estimate and then incorporate the ICC.

Define the variance of the treatment effect estimate in a CRE. Let $\hat{\tau}$ be the estimated treatment effect, and $Var(\hat{\tau})$ be its variance. In a CRE, the variance of the treatment effect estimate can be expressed as:

$$Var(\hat{\tau}) = \frac{\sigma_\epsilon^2 + \sigma_u^2}{n_1 n_2}$$

where:

σ_ϵ^2 is the within-cluster variance σ_u^2 is the between-cluster variance n_1 is the number of individuals per cluster n_2 is the number of clusters

The ICC, denoted as ρ, is defined as the proportion of the total variance that is due to between-cluster variation:

$$\rho = \frac{\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2}$$

Rearrange the ICC formula to express the between-cluster variance in terms of the ICC and total variance:

$$\sigma_u^2 = \rho(\sigma_\epsilon^2 + \sigma_u^2)$$

Substitute the expression for σ_u^2 into the variance formula from Step 1:

$$Var(\hat{\tau}) = \frac{\sigma_\epsilon^2 + \rho(\sigma_\epsilon^2 + \sigma_u^2)}{n_1 n_2}$$

Simplify the numerator by factoring out $(\sigma_\epsilon^2 + \sigma_u^2)$:

$$Var(\hat{\tau}) = \frac{(\sigma_\epsilon^2 + \sigma_u^2)(1 - \rho + n_1\rho)}{n_1 n_2}$$

When $ho = 0$ (no ICC), the variance formula reduces to:

$$Var(\hat{\tau}) = \frac{\sigma_\epsilon^2 + \sigma_u^2}{n_1 n_2}$$

As ρ increases, the term $(1 - \rho + n_1\rho)$ in the numerator increases, resulting in a larger variance of the treatment effect estimate.

Thus, as the ICC (ρ) increases, the variance of the treatment effect estimate in a CRE also increases. The ICC inflates the variance by a factor of $(1-\rho+n_1\rho)$, which depends on the cluster size and the magnitude of the ICC. This demonstrates the importance of accounting

for the ICC in the design and analysis of cluster-randomized experiments to obtain accurate estimates of the treatment effect and its precision.

B.2 Binned Estimator

For each bin b, we calculate the treatment effect estimator $\hat{\tau}_b$ as the difference between the mean outcomes of the treated and control groups [Imbens and Rubin, 2015]:

$$\hat{\tau}_b = \bar{y}_{1b} - \bar{y}_{0b} \tag{B.2.1}$$

where $\bar{y}_{1b}$ and $\bar{y}_{0b}$ are the mean outcomes for the treated and control groups in bin b, respectively. These means are calculated as:

$$\bar{y}_{1b} = \frac{1}{n_{1b}} \sum_{i=1}^{n_{1b}} y_{1bi} \tag{B.2.2}$$

$$\bar{y}_{0b} = \frac{1}{n_{0b}} \sum_{i=1}^{n_{0b}} y_{0bi} \tag{B.2.3}$$

Here, n_{1b} and n_{0b} represent the number of observations in the treated and control groups for bin b, while y_{1bi} and y_{0bi} denote the i-th observation in the treated and control groups for bin b, respectively.

The variance of the bin-specific treatment effect estimator, v_b, is calculated as:

$$v_b = \mathrm{Var}(\hat{\tau}_b) = \mathrm{Var}(\bar{y}_{1b}) + \mathrm{Var}(\bar{y}_{0b}) \tag{B.2.4}$$

$$v_b = \frac{\hat{\sigma}^2_{1b}}{n_{1b}} + \frac{\hat{\sigma}^2_{0b}}{n_{0b}} \tag{B.2.5}$$

where $\hat{\sigma}^2_{1b}$ and $\hat{\sigma}^2_{0b}$ are the sample variances of the treated and control groups in bin b, computed as:

$$\hat{\sigma}_{1b}^2 = \frac{1}{n_{1b}-1}\sum_{i=1}^{n_{1b}}(y_{1bi} - \bar{y}_{1b})^2 \tag{B.2.6}$$

$$\hat{\sigma}_{0b}^2 = \frac{1}{n_{0b}-1}\sum_{i=1}^{n_{0b}}(y_{0bi} - \bar{y}_{0b})^2 \tag{B.2.7}$$

The combined treatment effect estimator, $\hat{\tau}$, is a weighted average of the bin-specific estimators:

$$\hat{\tau} = \sum_{b=1}^{B} w_b \hat{\tau}_b \tag{B.2.8}$$

The weights, w_b, are inversely proportional to the variances of the bin-specific estimators:

$$w_b = \frac{1/v_b}{\sum_{b=1}^{B} 1/v_b} \tag{B.2.9}$$

This weighting scheme ensures that bins with more precise estimates (i.e., lower variance) contribute more to the overall estimator.

The variance of the combined estimator can be derived using the formula for the variance of a weighted sum of independent random variables:

$$Var(\hat{\tau}) = \sum_{b=1}^{B} w_b^2 v_b = \frac{1}{\sum_{b=1}^{B} 1/v_b} \tag{B.2.10}$$

Consequently, the standard error of the combined treatment effect estimator is given by:

$$SE(\hat{\tau}) = \sqrt{Var(\hat{\tau})} = \sqrt{\frac{1}{\sum_{b=1}^{B} 1/v_b}} \tag{B.2.11}$$

B.3 Consistency of Binned Estimator

Under Assumptions 1-7 we show that the binned estimator:

$$\hat{\tau} = \sum_{b=1}^{B} w_b \hat{\tau}_b$$

is a consistent estimator of the true Average Treatment Effect (ATE) τ.

Assumption 1: *(Random Assignment at the Cluster Level within Bins)* Within each bin b, treatment is randomly assigned at the cluster level, and all units within a cluster receive the same treatment. Formally, for cluster g in bin b:

$$W_{gb} \perp \{Y_{igb}(0), Y_{igb}(1)\}_{i \in C_{gb}}$$

Assumption 2: *(Independence Between Clusters within Bins)* Potential outcomes are independent across clusters within each bin:

$$\{Y_{igb}(0), Y_{igb}(1)\}_{i \in C_{gb}} \perp \{Y_{kgb}(0), Y_{kgb}(1)\}_{k \in C_{g'b}} \quad \text{for all } g \neq g'$$

Assumption 3: *(Cluster Alignment with Bins)* Each cluster is contained entirely within a single bin:

$$\text{If } i \in C_{gb} \text{ then } i \notin C_{gb'} \text{ for all } b' \neq b$$

Assumption 4: *(Consistent Estimation of Cluster-Robust Variances)* The estimated cluster-robust variances of the bin-specific treatment effects converge to their true values:

$$\hat{v}_b \xrightarrow{p} v_b \quad \text{as} \quad n_b \to \infty$$

Assumption 5: *(Proportional Growth of Sample Sizes Across Bins)* The fraction of

observations in each bin converges to a positive limit:

$$\lim_{n\to\infty} \frac{n_b}{n} = \pi_b > 0 \quad \forall b \in \{1, \ldots, B\}$$

Assumption 6: *(Convergence of Weights)* The weights based on inverse variances converge to well-defined limits:

$$w_b \xrightarrow{p} \omega_b \quad \text{where} \quad \sum_{b=1}^{B} \omega_b = 1$$

Assumption 7: *(Treatment Effect Homogeneity)* The treatment effect is homogeneous across bins:

$$\tau_b = \tau \quad \forall b \in \{1, \ldots, B\}$$

Proof: Under Assumptions 1-3, within each bin b, $\hat{\tau}_b$ is unbiased and consistent for τ_b when using cluster-robust standard errors following [Arellano, 1987].

Since $\tau_b = \tau$ (Assumption 7), we have:

$$\hat{\tau}_b \xrightarrow{p} \tau \quad \text{as} \quad n_b \to \infty$$

By Assumption 6, as $n \to \infty$:

$$w_b \xrightarrow{p} \omega_b \quad \text{where} \quad \sum_{b=1}^{B} \omega_b = 1$$

The binned estimator is:

$$\hat{\tau} = \sum_{b=1}^{B} w_b \hat{\tau}_b$$

As $n \to \infty$:

$$w_b \xrightarrow{p} \omega_b \quad \text{and} \quad \hat{\tau}_b \xrightarrow{p} \tau$$

Therefore, by the Continuous Mapping Theorem:

$$\hat{\tau} = \sum_{b=1}^{B} w_b \hat{\tau}_b \xrightarrow{p} \sum_{b=1}^{B} \omega_b \tau = \tau \sum_{b=1}^{B} \omega_b = \tau$$

Thus, the binned estimator converges in probability to the true ATE.

B.4 Data Plots

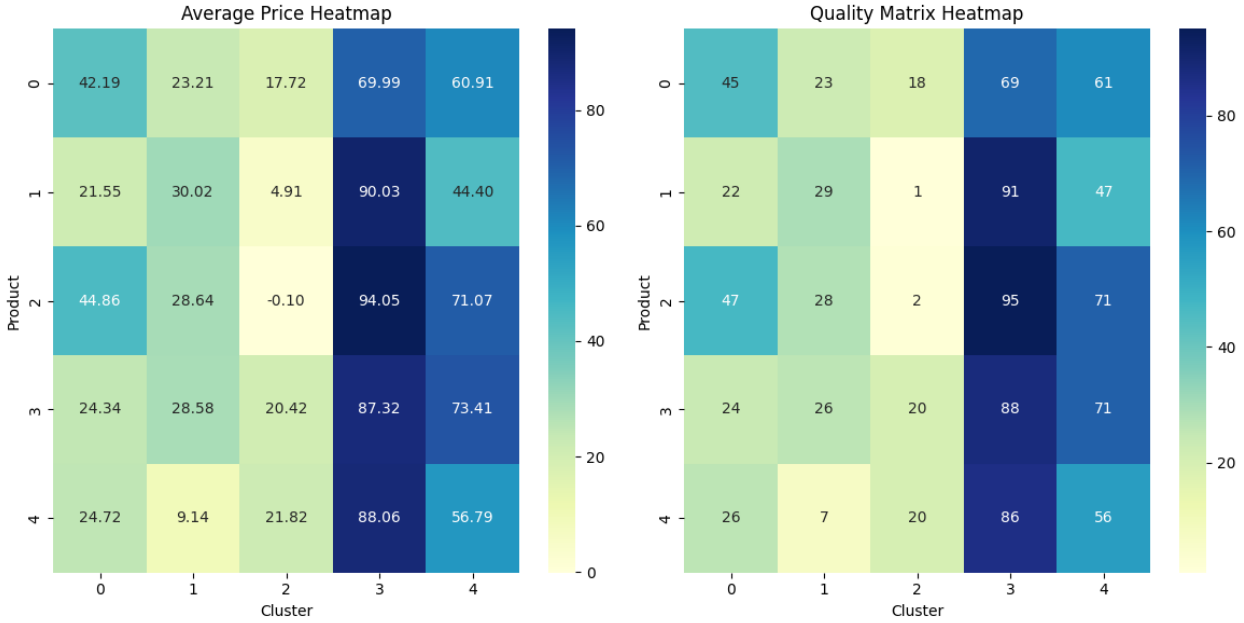
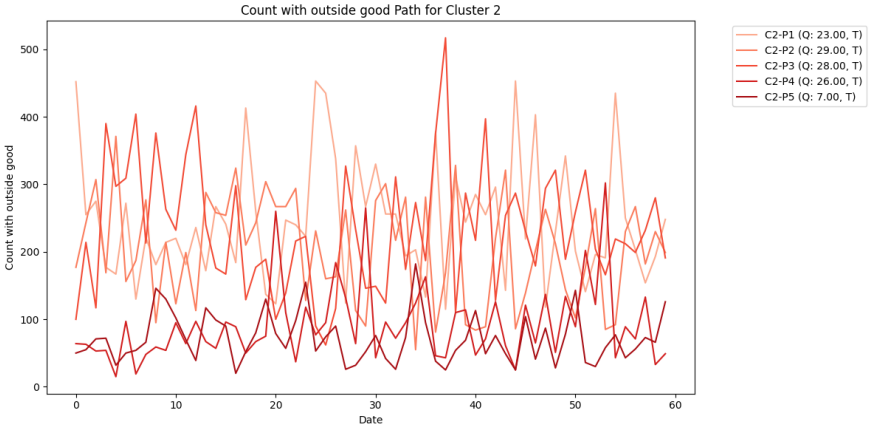
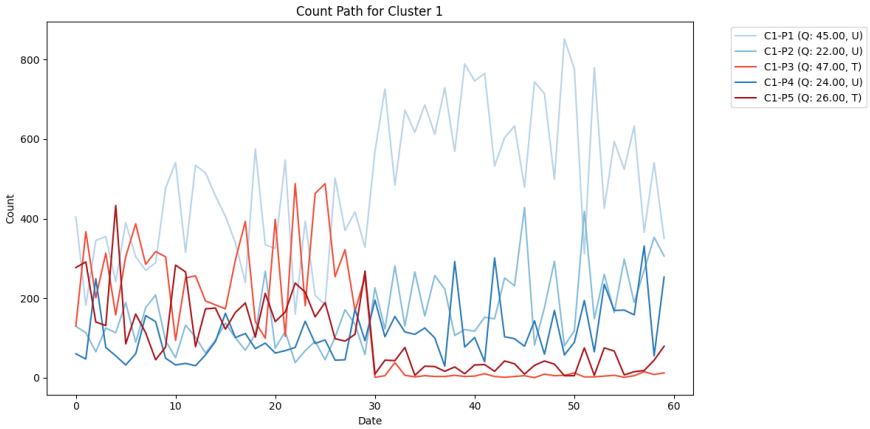


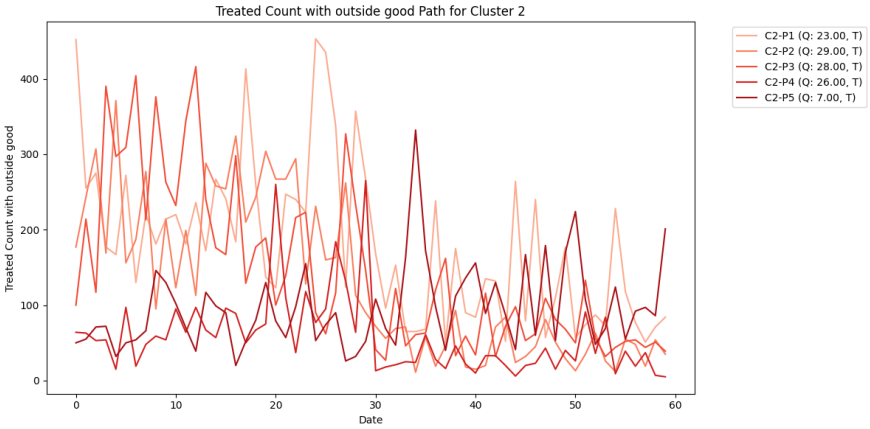
Figure B.1: Heatmaps of average price and average quality for the products in a sample of 5 clusters.



(a) Count paths for the products in Cluster 1, when no treatment is applied.



(b) Count paths for the products in Cluster 1, when subject-level randomization is applied. Red lines indicate the treated products, and blue lines indicate the products that are not treated.



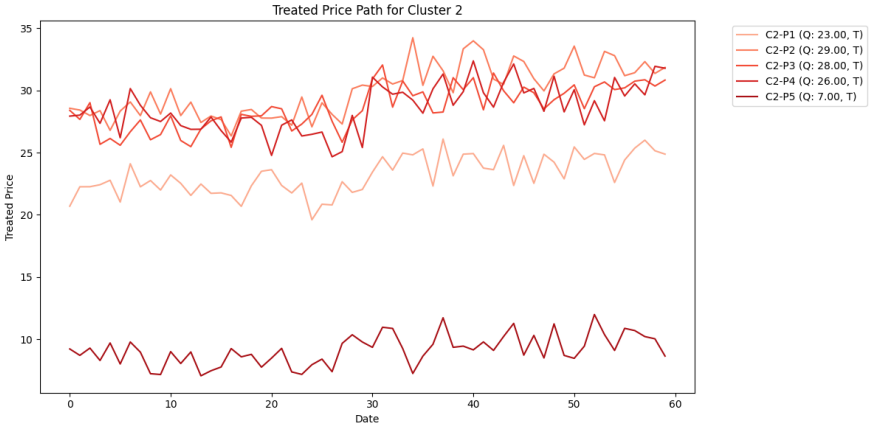
(c) Count paths for the products in Cluster 1, when cluster-level randomization is applied. Red lines indicate the treated products.

Figure B.2: Example count paths of the products.

Figure B.3: Example price paths of the products.

(a) Price paths for the products in Cluster 1, when no treatment is applied.

(b) Price paths for the products in Cluster 1, when subject-level randomization is applied. Red lines indicate the treated products, and blue lines indicate the products that are not treated.



(c) Price paths for the products in Cluster 1, when cluster-level randomization is applied. Red lines indicate the treated products.

B.5 Tables for Balance checks

Table I: Simple Subject Level Randomization: Comparison of Treatment and Control Groups

Variable	**T-mean**	**C-mean**	**t-stat**	**T-test P-value**	**KS-stat**	**KS-test P-value**
quality	52.07	53.16	-0.6690	0.5036	0.0287	0.9620
revenue	8817.39	8874.35	-0.1060	0.9156	0.0360	0.8222
profit	583.89	620.77	-0.8337	0.4047	0.0333	0.8850
count	178.70	172.90	0.7022	0.4827	0.0453	0.5577

Table II: Simple Cluster Level Randomization: Comparison of Treatment and Control Groups

Variable	**T-mean**	**C-mean**	**t-stat**	**T-test P-value**	**KS-stat**	**KS-test P-value**
quality	56.34	49.12	4.5506	0.0000	0.1008	0.0035
revenue	9476.21	8226.92	2.3792	0.0175	0.0928	0.0092
profit	670.84	541.19	3.0445	0.0024	0.1104	0.0010
count	174.98	175.46	-0.0601	0.9521	0.0384	0.7466

Table III: Simple Matched-Pair Cluster Randomization: Comparison of Treatment and Control Groups

Variable	**T-mean**	**C-mean**	**t-stat**	**T-test P-value**	**KS-stat**	**KS-test P-value**
quality	52.74	52.72	0.0150	0.9880	0.0192	0.9998
revenue	8858.29	8844.84	0.0256	0.9796	0.0352	0.8339
profit	603.04	608.99	-0.1392	0.8893	0.0480	0.4679
count	175.53	174.90	0.0788	0.9372	0.0304	0.9352

Table IV: Simple Stratified Cluster Randomization: Comparison of Treatment and Control Groups

Variable	**T-mean**	**C-mean**	**t-stat**	**T-test P-value**	**KS-stat**	**KS-test P-value**
quality	52.36	53.09	-0.4577	0.6472	0.0342	0.8396
revenue	8764.88	8936.87	-0.3268	0.7439	0.0335	0.8578
profit	625.12	587.21	0.8876	0.3749	0.0358	0.7981
count	173.33	177.07	-0.4675	0.6402	0.0559	0.2680

Table V: Simple Covariate Constrained Cluster Randomization: Comparison of Treatment and Control Groups

Variable	**T-mean**	**C-mean**	**t-stat**	**T-test P-value**	**KS-stat**	**KS-test P-value**
quality	53.43	52.02	0.8789	0.3796	0.0448	0.5576
revenue	8974.98	8728.15	0.4691	0.6391	0.0432	0.6046
profit	619.15	592.88	0.6148	0.5388	0.0512	0.3860
count	175.29	175.15	0.0172	0.9863	0.0352	0.8339

Appendix C

TWO STAGE TEXT REGRESSION USING TRANSFORMER-BASED ENCODINGS

C.1 Transformer Architecture

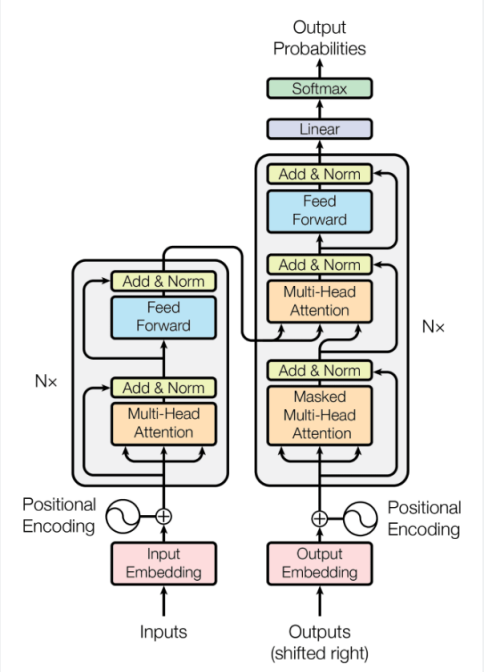


Figure C.1: Transformer neural network architecture.([Vaswani et al., 2017])

Transformers are highly modular neural network architectures that are very successful in sequence to sequence prediction. They are fitted in a self-supervised way, that is, the dependent variable that are predicted in the training stage is coming out of the text data itself. For example if the text data is "Love all, trust a few, do wrong to none." (Shakespeare), the input text - dependent variable pairs that are used in the training of the transformer evolve

in the following way:

- Input text: *love* - Dependent Variable: *all*
- Input text: *love all* - Dependent Variable: *trust*
- Input text: *love all trust* - Dependent Variable: *few*
- Input text: *love all trust few* - Dependent Variable: *do*, and so on

In the subsections below, we explain the parts of the transformer architechture that makes their empirical performance possible. Each of these ideas are coming from different papers in the neural network literature and used in the original transformer paper [Vaswani et al., 2017].

C.1.1 Positional Encodings

For the transformer to make use of the order of the sequence of words, positional encodings are inserted to the embeddings. Positional embeddings have the same dimension as the embeddings and they are summed together with the embeddings of the model so that the summation will include both the positional information and the position in the embedding space.

$$PE_{pos,2i} = sin(\frac{pos}{10000^{2i/d_{model}}})$$

$$PE_{pos,2i+1} = cos(\frac{pos}{10000^{2i/d_{model}}})$$

Intuitively each position is represented with different sinusoids (since they will have different wavelengths), and since the sine and cosine functions have values in $[-1, 1]$, their values are kept in a normalized range.

Note that [Vafa et al.,] uses a similar trick to incorporate the covariates X_i to the embeddings when they use the transformer architecture to predict the next occupation of an individual, using massive online resume data.

C.1.2 Self-Attention Mechanism

Attention is a communication mechanism which is a block in the neural network architecture that is created to allow that different words can attend to different parts of the past sentence when predicting the next word. The way that is implemented is as follows: Each word(or token) has a Key, Query, Value triplets as parameters of the model. Queries of the each token represents what that token is looking for, Keys represent the information that token has. Then, the dot product of Key and Query vectors are computed, scaled and converted to a weight matrix. Finally, the weight vector is multiplied with the value vector to compute the output of the attention block.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

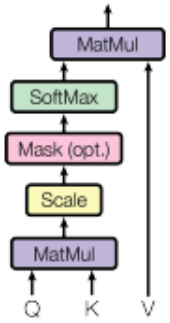


Figure C.2: Scaled dot-product attention.([Vaswani et al., 2017])

C.1.3 Skip Connections

Empirically, when the depth of a neural network is large, it is harder to optimize. In [He et al., 2016], it is shown with comprehensive emprical evidence that when skip connections are added to the architecture, it is significantly easier to optimize the parameters of the neural network. The idea is to sum the outputs of the initial layers of the neural network with the outpput layers so that during the optimization the gradients are distributed equally

among the arms of the summation and hence during the optimization all the information from the gradient of the objective function goes back to the initial layers hence somewhat reducing the depth of the neural network.

C.1.4 Layer Normalization

Similar to skip connections, Layer Norm is helpful for optimizing very deep neural networks, first introduced in [Ba et al., 2016]. For every observation, with say dimension P, compute the average and standard deviation of all of the covariates $\mu_i = \sum_{p=1}^{P} x_{ip}$ and normalize each with the computed average and standard deviation.

C.1.5 Dropout

Dropout is a regularization technique for training very large nerual networks, introduced in [Srivastava et al., 2014]. It works by dropping some random set of the connections of the neural networks to 0 during training time, hence prevents overfitting. It resembles the regularization scheme used in random forests which assume sparsity of the covariates. In random forests, each tree is estimated using a random subset of the covariates and in the final model all trees are averaged. In dropout, during the optimization some random sets of parameters are dropped to 0 and hence we can think of the final model as an ensemble of less dense neural networks.

C.2 Additional Numerical Results and Hyper-parameters

C.2.1 3-Class Classification Results

Table I: Comparison of 3-Class Classification Metrics for BERT MEAN Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
	FEMALE	0.71	0.34	0.46
	MALE	0.61	0.29	0.40
Random Forest	NEUTRAL	0.60	0.91	0.72
hyperparameters: Table V	**Accuracy**		0.62	
	Macro Avg	0.64	0.52	0.53
	Weighted Avg	0.63	0.62	0.58
	FEMALE	0.67	0.46	0.54
	MALE	0.59	0.41	0.48
LightGBM	NEUTRAL	0.66	0.86	0.74
hyperparameters: Table V	**Accuracy**		0.65	
	Macro Avg	0.64	0.58	0.59
	Weighted Avg	0.64	0.65	0.63
	FEMALE	0.67	0.44	0.53
	MALE	0.59	0.40	0.48
XGBoost	NEUTRAL	0.65	0.86	0.74
hyperparameters: Table V	**Accuracy**		0.64	
	Macro Avg	0.64	0.57	0.59
	Weighted Avg	0.64	0.64	0.63
	FEMALE	0.67	0.58	0.62
	MALE	0.61	0.54	0.57
MLP	NEUTRAL	0.73	0.82	0.77
hyperparameters: Table VI	**Accuracy**		0.69	
	Macro Avg	0.67	0.65	0.66
	Weighted Avg	0.69	0.69	0.69
	FEMALE	0.62	0.55	0.58
	MALE	0.59	0.48	0.53
Lasso Logistic	NEUTRAL	0.70	0.81	0.75
$\lambda = 0.001$	**Accuracy**		0.66	
	Macro Avg	0.64	0.61	0.62
	Weighted Avg	0.65	0.66	0.65

Table II: Comparison of 3-Class Classification Metrics for ADA v3 Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
	FEMALE	0.76	0.41	0.53
	MALE	0.64	0.29	0.40
Random Forest	NEUTRAL	0.62	0.92	0.74
hyperparameters: Table V	**Accuracy**		0.64	
	Macro Avg	0.67	0.54	0.56
	Weighted Avg	0.66	0.64	0.61
	FEMALE	0.71	0.51	0.60
	MALE	0.61	0.44	0.51
LightGBM	NEUTRAL	0.68	0.86	0.76
hyperparameters: Table V	**Accuracy**		0.67	
	Macro Avg	0.67	0.61	0.62
	Weighted Avg	0.67	0.67	0.66
	FEMALE	0.71	0.50	0.59
	MALE	0.61	0.43	0.51
XGBoost	NEUTRAL	0.67	0.87	0.76
hyperparameters: Table V	**Accuracy**		0.67	
	Macro Avg	0.67	0.60	0.62
	Weighted Avg	0.67	0.67	0.66
	FEMALE	0.72	0.61	0.66
	MALE	0.63	0.58	0.61
MLP	NEUTRAL	0.75	0.83	0.79
hyperparameters: Table VI	**Accuracy**		0.71	
	Macro Avg	0.70	0.67	0.68
	Weighted Avg	0.71	0.71	0.71
	FEMALE	0.66	0.59	0.63
	MALE	0.59	0.53	0.56
Lasso Logistic	NEUTRAL	0.73	0.80	0.77
$\lambda = 0.001$	**Accuracy**		0.69	
	Macro Avg	0.66	0.64	0.65
	Weighted Avg	0.68	0.69	0.68

Table III: Comparison of 3-Class Classification Metrics for BERT CLS Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
	FEMALE	0.69	0.26	0.38
	MALE	0.62	0.29	0.39
Random Forest	NEUTRAL	0.58	0.92	0.71
hyperparameters: Table V	**Accuracy**		0.60	
	Macro Avg	0.63	0.49	0.50
	Weighted Avg	0.62	0.60	0.55
	FEMALE	0.66	0.41	0.50
	MALE	0.59	0.40	0.48
LightGBM	NEUTRAL	0.64	0.86	0.74
hyperparameters: Table V	**Accuracy**		0.64	
	Macro Avg	0.63	0.56	0.57
	Weighted Avg	0.63	0.64	0.61
	FEMALE	0.66	0.39	0.49
	MALE	0.59	0.38	0.46
XGBoost	NEUTRAL	0.63	0.87	0.73
hyperparameters: Table V	**Accuracy**		0.63	
	Macro Avg	0.63	0.55	0.56
	Weighted Avg	0.63	0.63	0.61
	FEMALE	0.67	0.53	0.59
	MALE	0.59	0.57	0.58
MLP	NEUTRAL	0.73	0.82	0.77
hyperparameters: Table VI	**Accuracy**		0.68	
	Macro Avg	0.66	0.64	0.65
	Weighted Avg	0.68	0.68	0.68
	FEMALE	0.63	0.50	0.55
	MALE	0.55	0.49	0.52
Lasso Logistic	NEUTRAL	0.69	0.80	0.74
$\lambda = 0.001$	**Accuracy**		0.65	
	Macro Avg	0.62	0.60	0.61
	Weighted Avg	0.64	0.65	0.64

Table IV: Comparison of 3-Class Classification Metrics for SBERT Embeddings & Second-Stage Models

Model	Class	Precision	Recall	F1-Score
	FEMALE	0.71	0.34	0.46
	MALE	0.61	0.22	0.32
Random Forest	NEUTRAL	0.58	0.92	0.71
hyperparameters: Table V	**Accuracy**		0.60	
	Macro Avg	0.64	0.49	0.50
	Weighted Avg	0.62	0.60	0.56
	FEMALE	0.66	0.44	0.53
	MALE	0.56	0.37	0.45
LightGBM	NEUTRAL	0.64	0.85	0.73
hyperparameters: Table V	**Accuracy**		0.63	
	Macro Avg	0.62	0.55	0.57
	Weighted Avg	0.63	0.63	0.61
	FEMALE	0.66	0.44	0.52
	MALE	0.57	0.34	0.43
XGBoost	NEUTRAL	0.63	0.86	0.73
hyperparameters: Table V	**Accuracy**		0.63	
	Macro Avg	0.62	0.55	0.56
	Weighted Avg	0.62	0.63	0.60
	FEMALE	0.66	0.54	0.59
	MALE	0.55	0.55	0.55
MLP	NEUTRAL	0.72	0.78	0.75
hyperparameters: Table VI	**Accuracy**		0.67	
	Macro Avg	0.64	0.62	0.63
	Weighted Avg	0.66	0.67	0.66
	FEMALE	0.63	0.46	0.53
	MALE	0.53	0.45	0.49
Lasso Logistic	NEUTRAL	0.67	0.81	0.73
$\lambda = 0.001$	**Accuracy**		0.63	
	Macro Avg	0.61	0.57	0.58
	Weighted Avg	0.63	0.63	0.62

C.2.2 Hyper-parameters of second stage models

Table V: Comparison of Hyperparameters for BERT-based Models

Hyperparameter	**Description**	**Random Forest (hp-1)**	**LightGBM (hp-2)**	**XGBoost (hp-3)**
bootstrap	Bootstrap samples	True	-	-
ccp_alpha	Pruning parameter	0.0	-	-
class_weight	Class weights	balanced	-	-
criterion	Split quality	gini	-	-
max_depth	Max tree depth	None	-1	10
max_features	Max features for split	sqrt	-	-
max_leaf_nodes	Max leaf nodes	None	-	-
min_samples_leaf	Min samples at leaf	1	-	-
min_samples_split	Min samples for split	2	-	-
n_estimators	Number of trees	100	-	-
n_jobs	Parallel jobs	None	-	4
random_state	Random seed	42	-	42
verbose	Verbosity	0	-1	-
objective	Objective	-	binary	binary:logistic
metric	Metric	-	binary_logloss	logloss
learning_rate	Learning rate	-	0.05	0.05
num_leaves	Max leaves	-	31	-
feature_fraction	Feature fraction	-	0.9	-
bagging_fraction	Bagging fraction	-	0.8	-
bagging_freq	Bagging frequency	-	5	-
subsample	Data subsample	-	-	0.8
colsample_bytree	Column subsample	-	-	0.9
nthread	Threads used	-	-	4

Table VI: Hyperparameters for MLP Model (hp-4)

Hyperparameter	**Value**
Input Dimension	768
Hidden Layer 1	512 units, ReLU, Dropout(0.3)
Hidden Layer 2	256 units, ReLU, Dropout(0.3)
Hidden Layer 3	128 units, ReLU, Dropout(0.3)
Output Layer	2 units (FEMALE, MALE)
Dropout Rate	0.3
Batch Size	32
Learning Rate	0.001
Number of Epochs	20
Loss Function	CrossEntropyLoss
Optimizer	Adam
Device	GPU (if available)

C.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that represents each document as a mixture of latent topics, and each topic as a distribution over words. Given a corpus of n documents and a vocabulary of V unique words, LDA assumes the following generative process for each document:

C.3.1 Generative Process

For each document $s_i \in \{s_1, \ldots, s_n\}$, LDA assumes the following steps:

1. Draw a topic distribution θ_i for the document from a Dirichlet prior with hyperparameter α:

$$\theta_i \sim \text{Dirichlet}(\alpha)$$

   where θ_i is a K-dimensional vector, and K is the number of topics.

2. For each topic $k \in \{1, \ldots, K\}$, draw a word distribution ϕ_k from a Dirichlet prior with hyperparameter β:

$$\phi_k \sim \text{Dirichlet}(\beta)$$

where ϕ_k is a V-dimensional vector representing the word distribution for topic k.

3. For each word $w_{i,n}$ in document s_i, where $n \in \{1, \ldots, N_i\}$:

   (a) Sample a topic $z_{i,n}$ from the topic distribution θ_i:

   $$z_{i,n} \sim \text{Multinomial}(\theta_i)$$

   where $z_{i,n} \in \{1, \ldots, K\}$ indicates the topic assignment for the n-th word in document s_i.

   (b) Given the topic $z_{i,n}$, sample a word $w_{i,n}$ from the topic-specific word distribution $\phi_{z_{i,n}}$:

   $$w_{i,n} \sim \text{Multinomial}(\phi_{z_{i,n}})$$

C.3.2 Joint Distribution

The joint distribution of the observed words w, the topic assignments z, the document-level topic distributions θ_i, and the topic-level word distributions ϕ_k is given by:

$$P(w, z, \theta, \phi \mid \alpha, \eta) = \prod_{i=1}^{n} \left[P(\theta_i \mid \alpha) \prod_{n=1}^{N_i} P(z_{i,n} \mid \theta_i) P(w_{i,n} \mid z_{i,n}, \phi) \right] \prod_{k=1}^{K} P(\phi_k \mid \eta)$$

where:

- $P(\theta_i \mid \alpha) = \text{Dirichlet}(\theta_i \mid \alpha)$ is the prior distribution of topics for document s_i.
- $P(\phi_k \mid \eta) = \text{Dirichlet}(\phi_k \mid \eta)$ is the prior distribution over words for topic k.
- $P(z_{i,n} \mid \theta_i) = \text{Multinomial}(\theta_i)$ is the probability of assigning the n-th word in document s_i to a topic $z_{i,n}$.
- $P(w_{i,n} \mid z_{i,n}, \phi) = \text{Multinomial}(\phi_{z_{i,n}})$ is the probability of generating word $w_{i,n}$ from the word distribution $\phi_{z_{i,n}}$ corresponding to topic $z_{i,n}$.

The primary goal of LDA is to infer the hidden topic structure from a set of documents, which involves estimating the posterior distribution of the latent variables given the observed words:

$$P(\theta, \phi, z \mid w, \alpha, \eta) = \frac{P(w, z, \theta, \phi \mid \alpha, \eta)}{P(w \mid \alpha, \eta)}$$

Since exact inference of this posterior distribution is intractable due to the integrals involved, approximate inference techniques such as Variational Inference or Gibbs Sampling are typically used. The parameters α and η are usually estimated using a maximum likelihood estimation procedure over the corpus. The topic distributions θ_i and word distributions ϕ_k are latent variables, and their estimates can be obtained through posterior inference after training the model on the corpus.

Once the model has been fit, LDA provides the distribution of topics for each document, θ_i, which indicates the mixture of topics present in document s_i, the distribution of words for each topic, ϕ_k, which indicates the most likely words associated with topic k and the topic assignment $z_{i,n}$ for each word in the document, which can be used to identify the dominant topic of each word or document.