Payment Instrument Choice with Scanner Data: An MM algorithm for Fixed Effects in Non-Linear Models*

Mingli Chen
University of Warwick

Marc Rysman
Boston University

Shuang Wang
Boston University

Krzysztof Wozniak
Federal Reserve Bank

November 30, 2020

Click here to download the most recent version

1 Introduction

Over the past several decades, the US payments system has shifted from paper payment instruments, namely cash and check, to digital instruments, such as debit cards and credit cards. This shift is important because digital payments are typically regarded as superior in most dimensions: they are faster and cheaper to process, and they are easier to track and less subject to crime. The shift to digital payments is far from complete however, as cash and check still play a large role in the economy, particularly in some sectors.

This paper studies the determinants of payment choice over short and long horizons. Over short horizons, for instance within a quarter or a month, we focus on the transaction size as an important determinant. Transaction size has been central to the discussion of payment choice. Typically, consumers switch away from cash as the transaction size becomes larger. Previous research has studied the effect of transaction size on payment choice by using scanner data drawn from retailers. Leading examples are Klee (2008) and Wang & Wolman (2016). However, these papers rely on data sets that do not allow researchers to track individuals over time, and thus the measure they provide is a conflation of the within and between effect. For instance, we do not know whether households switch to card as transaction size gets larger, or households never switch but households that use cards more often tend to have higher transaction sizes on average. A central goal of this paper is to separate these effects.

We also study the long-term evolution of preferences for payment mechanisms. In particular, the observed increase in card usage may be a result of changes in household preferences in favor of cards. However, alternative explanations are that there are shifts in the composition of transactions or transaction sizes. For instance, if older households prefer cash and check and also experience decreases in transactions, while younger households prefer card and experience increases in transactions, we will observe an aggregate increase in card usage although no household has experienced a change in preferences for card relative to cash and check. Our paper decomposes these compositional changes from changes in preferences over a five-year period.

*Researchers own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

This paper leverages consumer scanner data set to obtain transaction-level data on payment choice. Nielsen maintains a panel of households that tracks in great detail their purchase choices of food and non-food items, across all retail outlets in all U.S markets (except Alaska and Hawaii). These types of data are common for marketing studies. In addition, Nielsen tracks the method of payment of each choice. We access the data through the Kilts Center of the University of Chicago, which recently made the payment choice data available.[1] To our knowledge, no previous academic study has used such data to study payment choice.

Thus, our goal is to estimate a within-household effect of transaction size on payment choice using panel data. To do so, we estimate a multinomial discrete choice model with household-choice fixed effects. As we are interested in how preferences change over time, we allow the household-choice fixed effects to vary by quarter. With three choices, 10,000 households, and 20 quarters of data, this problem creates a significant numerical challenge. Standard procedures for estimating multinomial logit models fail with this many parameters. An important contribution of our paper is to introduce a new method to address this numerical problem. We believe our solution is applicable in a wide variety of setting where a researcher wishes to estimate fixed effects in a non-linear models.

We rely on the Minorization-Maximization (MM) algorithm, which has been developed in the statistics literature, as in Hunter & Lange (2004); Lange (2016), but has almost no applications in econometrics. The MM algorithm can be seen as a generalization of the EM algorithm. They are identical in the case of the binary probit, but as we show below, we require the MM algorithm to achieve computational benefits in the binary or multinomial logit models. We utilize the MM algorithm to linearize the logit model so that we can apply linear techniques, such as demeaning, to the fixed effects estimation. Sequential fixed effects estimation and minorization allows us to find numerically identical estimates to maximum likelihood at a tiny fraction of the computational and memory costs that the dummy variables estimator would entail.

Because we estimate many fixed effects in a panel setting, we face the incidental parameters problem. As in several previous papers, we address the incidental parameters problem with ex-post bias reduction via the jackknife following Dhaene & Jochmans (2015). We find that accounting for household-choice fixed effects and household-choice-quarter fixed effects substantially affects the estimated importance of transaction size on payment method choice. For instance, we find that going from the 1st quantile of the empirical distribution of transaction size, \$11.94, to the 3rd quantile, \$57.06 leads to a 18.6 percentage point increase of probability of using a card. In order to simulate previous research, we also estimate our model without fixed effects and show that it leads a 11.2% larger estimated effect. That is, accounting for household heterogeneity substantially reduces the estimated impact of transaction size on payment choice.

Further, we find that over the five-year period of the data, aggregate value-weighted card usage increases by 9.73 percentage points. We provide a decomposition of this change into changes in household preferences (i.e., changes in household-choice-quarter fixed effects), and changes in the number and value of transactions, and entry and exit of households from the sample. We find that only about a third of the change is due to changes in household preferences. The rest is due to compositional changes in who makes payments and of what value. To the extent that household preferences are changing relatively slowly, public policy efforts to move consumers to electronic payments may be of relatively little value.

Overall, our paper makes several contributions. We provide an attractive new approach to estimate multinomial discrete choice models with fixed effects. Within the MM literature, we provide a new formalization of the MM algorithm and a new minorization for the multinomial logit model that could likely be extended to a number of linear-index likelihood models. We introduce consumer scanner data, typically used for studying purchases of groceries, as a powerful tool for studying payment choice. We present new results about the importance of transaction size in determining payment choice, for the first time accounting for persistent unobserved household heterogeneity. We decompose long-term trends in payment choice and show the relatively limited role that changes in preferences have in driving these trends.

[1]The Kilts Center requested payment choice data from Nielsen in part based on our request.

2 Literature Review

A number of studies aim to identify the determinants of payment choice. However, doing so is often hampered by data constraints. It is difficult to track the payments of individual households, particularly with regard to cash. One method for tracking payment choice is to survey consumers retrospectively such as in Schuh & Stavins (2010) and Koulayev, Rysman, Schuh & Stavins (2016), which use a survey that asks consumers about payment use over the previous month, for instance, the number of times a household used a credit card in the last 30 days. However, this method makes it difficult to study the determinants of each individual choice, or why choice varies across shopping trips. Another method is to ask survey participants to fill out a diary of payment behavior, such as in Rysman (2007), Arango, Huynh & Sabetti (2015) and Wakamori & Welte (2017). This is an important contribution, although Jonker & Kosse (2009) raises questions about how accurate these surveys are, showing that the daily number of transactions in seven-day surveys is significantly less than in one-day surveys, suggesting a form of "diary fatigue." A solution to this problem is to obtain data directly from consumer bank accounts so consumers are passive, such as in White (1975), Stango & Zinman (2014) and Dutkowsky & Fusaro (2011). However, these typically provide no information on how the consumer uses cash, and consumers may use multiple accounts for transactions, some of which may not show up in the available transaction record.

Scanner data has important advantages over these alternatives. We observe individual household decisions continuously for a period of three years, something that no existing diary data set can come close to matching. We observe important demographics such as age, household size and income. Our data has important limitations. We do not observe every transaction a household makes. The data set is probably most complete with regard to grocery payments, and groceries are an important touchpoint for payment choice, and have been a focus of the payments industry. Also, another issue is that the method that Nielsen uses for tracking payments is not perfect for our purposes, as we essentially cannot distinguish between debit and credit use. But importantly, we can distinguish between cash, check and card, and we observe transaction size, which is the focus of the paper. We discuss further limitations below.

A closely related paper is Klee (2008). Klee also uses scanner data from grocery purchases to study payment choice. Her data set is drawn from the cash register of a grocery chain. As a result, she cannot observe the identities of the purchasers, and thus cannot track consumers over time. She accounts for consumer demographics by using census data on the neighborhoods of the stores. This contrasts with our paper, where we observe consumer demographics directly and can account for unobserved heterogeneity using panel techniques such as fixed effects. In addition, our study covers packaged food shopping from a wide array of retailing channels, not just a single store. Like us, Klee cannot distinguish between debit and credit, although she can distinguish signature and PIN-based card transactions. Wang & Wolman (2016) follows a similar approach. Most of the papers we discuss here rely on data sets that cover relatively short time periods. We are not aware of another paper that attempts to decompose long-term changes in card use similar to what we do here.

Our paper is also a contribution to estimating fixed effects in non-linear models. Several other papers precede us in this regard. A classic contribution is Chamberlain's conditional logit model (Chamberlain, 1980). Typical implementations handle only the binary outcome case, and extending the model to multinomial outcomes creates significant combinatoric complexities. Furthermore, the model does not naturally deliver estimates for the fixed effects, which our approach does.

Like us, several papers advocate for a computational approach to estimating the fixed effects model and then using the jackknife for bias correction. Hospido (2012) and D'Haultfœuille & Iaria (2016) introduce efficient methods for computing the dummy variables model and, like us, rely on ex-post bias correction. Hospido (2012) exploits the sparsity of the fixed effects and D'Haultfœuille & Iaria (2016) rely on simulation (which introduces integration error) of the choice set to cheaply compute the Hessian of the objective function. However, in our application, the Hessian is larger than can be addressed in many standard computer set-ups.

Stammann, Heiß & McFadden (2016) is perhaps closest to us in that they advocate for iterative demeaning to obtain estimates, and then ex-post bias correction, in their case based on Hahn & Newey (2004). Several techniques rely on concentrating out fixed effects so the researcher can maximize only over the remaining parameters, such as Hinz, Hudle & Wanner (2019) and Stamann (2018). Our understanding is that all three of these approaches have been developed only for binary outcomes, and do not easily expand to mulitnomial settings. Another approach relies on differencing out fixed effects in way that leads to estimation with moment inequalities, such as Ho & Pakes (2014) and Shum, Song & Shi (2018). Our approach differs in that it generates point identification and delivers estimates of the fixed effects, and we view our approach as computationally less challenging than estimation with moment inequalities.

The Minorization-Maximation algorithm, which is sometimes called the Majorization-Minorization algorithm but is the MM algorithm in either case, has a long history in statistics dating about to the time of the introduction of the EM algorithm. The EM algorithm can be regarded as a special case of the MM algorithm. In general, the MM algorithm expands the set of functions that can be used in the E-step of the EM algorithm, and has appeared under many names in different papers, often depending on what function was used. Böhning & Lindsay (1988) is an important early citation and Hunter & Lange (2004) and Lange (2016) provides a helpful overview and history. We are aware of only one paper in the econometrics literature: James (2017) shows that the MM algorithm can be advantageous in the context of the mixed multinomial logit model of McFadden & Train (2000) but does not discuss the application to fixed effects or dynamic models, or provide an application.

Our paper is close to that of Chen (2019). She uses the EM algorithm in the context of the binary probit to estimate a model with interactive fixed effects. She utilizes the EM algorithm to obtain a linear form of the model and then applies known techniques for handling interactive fixed effects in the linear case, and she uses ex-post bias correction, in her case a known analytic form. In fact, her implementation of the EM algorithm bares a resemblance to the MM algorithm. She does not consider multinomial models. Following her ideas, our model could be extended to handle interactive fixed effects in a multinomial logit model.

3 The Minorization - Maximization (MM) Estimation Procedure

The general idea of this iterative estimation procedure is that in each interation, we (1) construct a simpler concave surrogate function that minorizes the complicated log-likelihood function, i.e. the surrogate function is less than the log-likelihood function everywhere but equal to it at the current best guess of the parameters; (2) maximize the surrogate function instead of the log-likelihood function. The ascent of the log-likelihood is guaranteed by the property of minorization. By alternating between these steps of minorization and maximization, the MM algorithm finds the parameters that maximize the original log likelihood function.

Before diving into multnimonial logit model, we first discuss the definition of the MM algorithm in general and some conditions required for convergence.

3.1 The Transfer Minorization

Relative to the standard mathematical definition of a minorizing function, we add an extra condition that makes the function suitable for use in a maximization problem. As such, we refer to our minorization as a *transfer minorization.* Our name is based on the terminology of Lange, Hunter & Yang (2000), which refers to the minorization as the *transfer function.* That is, we transfer optimization from the function of interest to the minorization of this function.

Definition 1. *Suppose $\mathcal{L}$ is a real-valued function on Ω that is twice differentiable and S is a*

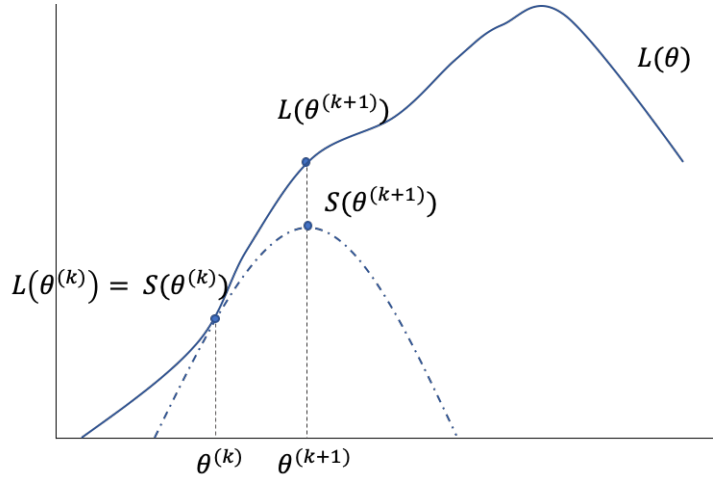

Figure 1: MM algorithm

real-valued function on $\Omega \otimes \Omega$*, we say that* S *is a* transfer minorization *of* $\mathcal{L}$ *if:*

(a) $S(\boldsymbol{\theta}; \boldsymbol{\theta}') \leq \mathcal{L}(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$*;*

(b) $S(\boldsymbol{\theta}'; \boldsymbol{\theta}') = \mathcal{L}(\boldsymbol{\theta}')$ *for all* $\boldsymbol{\theta}'$*;*

(c) $\nabla^{20} S(\boldsymbol{\theta}; \boldsymbol{\theta}')$ *exists, and is negative definite at* $\boldsymbol{\theta}$*.*

where $\nabla^{mn} S(\boldsymbol{\theta}; \boldsymbol{\theta}')$ *is the* m^{th} *order derivative w.r.t.* $\boldsymbol{\theta}$ *and* n^{th} *order derivate w.r.t to* $\boldsymbol{\theta}'$*.*

Analogously, S is a *transfer majorization* of $\mathcal{L}$ if $-S$ is a transfer minorization of $-\mathcal{L}$. The first two conditions of Definition 1 are from de Leeuw & Lange (2009) and are standard for defining a minorization. The third condition ensures that the minorization is well-behaved around the focal point. Arguably, we could use a less strict condition, such as that the minorization has the same sign as $\mathcal{L}$ in some region around $\boldsymbol{\theta}$, but in practice, we are not aware of any implementations of the MM algorithm that do not satisfy the third condition. As shown in de Leeuw & Lange (2009), an implication of Definition 1 is:

Corallary 1. *If* S *is a transfer minorization of* $\mathcal{L}$*, then for all* $\boldsymbol{\theta}$*:*

$$\nabla^{10} S(\boldsymbol{\theta}; \boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$$

This is basically the necessary condition that $\boldsymbol{\theta}$ minimizes the distance between $S(\cdot; \boldsymbol{\theta})$ and $\mathcal{L}(\cdot)$.

Intuitively, the idea is that when faced with a likelihood function $\mathcal{L}$ that is difficult to maximize, we instead maximize some other function S. The function S is chosen to be easy to maximize so that its maximand is always closer to a local optima than the current guess. Figure 1 provides an example. In the figure, we would like to find the optimum of $\mathcal{L}$ and we start with a guess $\theta^{(k)}$. Rather than seek to optimize $\mathcal{L}$ directly, we construct a transfer minorization $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$. As a minorization, it is always below $\mathcal{L}$ and is equal to $\mathcal{L}$ at $S(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})$. As a transfer minorization, S is well-behaved around $\boldsymbol{\theta}^{(k)}$, i.e. it is differentiable and concave. It is optimized at the point $\boldsymbol{\theta}^{(k+1)}$. At this point, we will construct a new transfer minorization $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k+1)})$ (not shown in the figure). Iterative application of this process leads to the maximum of $\mathcal{L}$, as shown in the next proposition. First, we define the MM algorithm:

Definition 2. *Let* $\boldsymbol{\theta}^{(0)}$ *be the initial guess of* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}^{(k)}$ *be the guess after* k *cycles of the algorithm. The* Minorization-Maximization (MM) Algorithm iteratively applies the following two-step procedure:

1. Minorization step: Compute $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.

2. Maximization step: Choose $\boldsymbol{\theta}^{(k+1)}$ to be a value of $\boldsymbol{\theta} \in R^p$ that maximizes $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.

At each step, let $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k+1)}$. Repeat steps 1 and 2 until $\boldsymbol{\theta}^{(k)}$ converges.

To show convergence, we generalize the Theorem 4 in Dempster, Laird & Rubin (1977) from the EM algorithm context to MM algorithm. In Theorem 1, we list conditions for the sequence $\boldsymbol{\theta}^{(k)}, k = 0, 1, 2, ...$ to converge to a point where $\nabla \mathcal{L}(\cdot) = 0$ in the context of maximum likelihood estimation (MLE). Specifically, a likelihood function $f(x; \boldsymbol{\theta})$ is the density of the true DGP $f(x; \boldsymbol{\theta}_0)$ with the true parameter vector $\boldsymbol{\theta}_0$ replaced with its hypothetic value $\boldsymbol{\theta}$.[2]

Theorem 1. *Suppose that $\mathcal{L}(\boldsymbol{\theta})$ is the log likelihood function and $S(\boldsymbol{\theta}; \boldsymbol{\theta}')$ is a minorization of $\mathcal{L}(\boldsymbol{\theta})$ for maximization, and $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}), k = 0, 1, 2, ...$ is an instance of an MM algorithm, then:*

1. *$\boldsymbol{\theta}^{(k)}$ converges to a $\boldsymbol{\theta}^*$ in the closure of Ω.*

2. *$\nabla \mathcal{L}(\boldsymbol{\theta}^*) = 0$, $\nabla^{20} S(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*)$ is negative definite with eigenvalues bounded away from zero.*

Proof. See Appendix A.1.[3] □

Thus, any function that satisfies Definition 1 for some function $\mathcal{L}$ can be used in the MM algorithm to find an optimum to $\mathcal{L}$.[4] This approach allows substantial freedom in selecting the transfer minorization. From the perspective of the MM algorithm, the EM algorithm is a special case and it works because the conditional expectation of $\mathcal{L}$ used in the EM algorithm is a transfer minorization. If the E-step of the EM algorithm does not deliver a function that is easy to optimize, as in our case, the researcher is free to use some other minorizing function. Hunter & Lange (2004) discuss several approaches. In the next section, we focus on a Taylor expansion, which delivers a least-squares optimization problem in our context.

4 A Minorization for the Multinomial Logit

In this section, we first present the multinomial logit model and then develop a transfer minorization and discuss our estimation method.

[2]More formally, we consider the log likelihood function as defined in (Hayashi, 2000, p.448). In particular, let $\{\boldsymbol{x}_n\}$ be an *i.i.d.* sequence where the density of $\boldsymbol{x}_n$ can be indexed by a finite-dimensional vector $\boldsymbol{\theta}_0$: $f(\boldsymbol{x}_n; \boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0 \in \Omega$. Because $\{\boldsymbol{x}_n\}$ is independently distributed, the joint density of the data $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ at a hypothetical value $\boldsymbol{\theta}$ is

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N; \boldsymbol{\theta}) = \prod_{n=1}^{N} f(\boldsymbol{x}_t; \boldsymbol{\theta}).$$

This density, viewed as a function of $\boldsymbol{\theta}$ is called the *likelihood function*. The maximum likelihood (ML) estimator of $\boldsymbol{\theta}_0$ is the $\boldsymbol{\theta}$ that maximizes the likelihood function. Because the log transformation is a monotone transformation, maximizing the likelihood function is equivalent to maximizing the log likelihood function.

$$\mathcal{L}(\boldsymbol{\theta}) = \ln\ f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N; \boldsymbol{\theta}) = \ln \left[\prod_{n=1}^{N} f(\boldsymbol{x}_t; \boldsymbol{\theta}) \right].$$

[3]We provide Theorem 1 and the rest of the formalism in this section because we could not find the mathematical statement we required in the existing literature. But to be clear, our approach relies heavily on the cited literature and for sure, the existing literature seems to operate as if it is well-known that the MM algorithm converges to a local optima.

[4]The papers that offer the closest version of what we present in this section define a minorization based only on parts (a) and (b) of Definition 1 and then include something like part (c) in the supposition of the analog to Theorem 1. For example, see Böhning & Lindsay (1988). We prefer to have all of the requirements in Definition 1 so we can simply check if a candidate function satisfies the definition to know that it can be used in the MM algorithm. Because our definition differs in this way from the previous research, we coin a new name for our version: the transfer minorization. But to be clear, we rely closely on existing contributions in our approach.

4.1 Model

In the multinomial logit model, an agent makes a discrete choice among several options, each of which draws an Extreme Value error. The model is distinguished by a closed-form logistic function for the probability of each choice. In our presentation, we emphasize a fixed effect that varies by household, choice and quarter.

We observe N consumers make a discrete choice among J products in each of T time periods. Consumer i in period t that chooses product j obtains utility u_{ijt}. Utility is defined as:

$$u_{ijt} = x_{it}\beta_j + \xi_{ijq(t)} + \varepsilon_{ijt}, \tag{1}$$

where $q(t)$ is the quarter of the year that t falls in where $q \in \{1, \ldots, Q\}$, x_{it} is a vector observable characteristics that varies by consumer and time, ξ_{ijq} is a consumer-product-quarter fixed effect and ε_{ijt} is a scalar *i.i.d* idiosyncratic shock distributed according to a Type I Extreme Value distribution. The variable y_{ijt} is a binary indicator for the product that consumer i chooses in t, where:

$$y_{ijt} = \mathbb{1}[u_{ijt} \geq u_{ikt}, \forall k \neq j], \ \forall i, t.$$

The parameter vectors β_j and ξ_{ijq} are to be estimated. We collect these parameters as $\boldsymbol{\theta} = (\{\beta_j\}_{j=1,\ldots,J}, \{\xi_{ijq}\}_{i=1,\ldots,N;j=1,\ldots,J;q=1,\ldots,Q})$. Then the log-likelihood function can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i,t} \log \ l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}; \boldsymbol{y}_{it}) \tag{2}$$

where $x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)} = (x_{it}\beta_1 + \xi_{i1q(t)}, x_{it}\beta_2 + \xi_{i2q(t)}, \ldots, x_{it}\beta_J + \xi_{iJq(t)})$, $\boldsymbol{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{iJt})$ and:

$$\begin{aligned} l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}; \boldsymbol{y}_{it}) &= \textstyle\prod_{j=1}^{J} \left(p_{ijt}(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}) \right)^{y_{ijt}}, \\ p_{ijt}(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}) &= \frac{\exp(x_{it}\beta_j + \xi_{ijq(t)})}{\sum_{k=1}^{J} \exp(x_{it}\beta_k + \xi_{ikq(t)})}. \end{aligned}$$

In practice, we must normalize the mean utility of one of the choices to be zero, as is standard in the multinomial logit.

4.2 The transfer minorization

This subsection derives a function S that satisfies the conditions of Definition 1 to be a transfer minorization of $\mathcal{L}$.

Theorem 2. *Let $\mathcal{L}(\boldsymbol{\theta})$ be the log likelihood function for a multinomial logit model defined as in Eq.(2). Let $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ be defined as:*

$$\begin{aligned} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)}) &+ \frac{1}{2} \sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it})^2 \\ &- \frac{1}{2} \sum_{i,j,t} \left(x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it}) - x_{it}\beta_j - \xi_{ijq(t)} \right)^2, \end{aligned} \tag{3}$$

where

$$\begin{aligned} h_j(\boldsymbol{\phi}_{it}^{(k)}; \boldsymbol{y}_{it}) &= \frac{\partial \log l(\boldsymbol{\phi}_{it}^{(k)}; \boldsymbol{y}_{it})}{\partial \phi_{ijt}^{(k)}} = y_{ijt} - p_{ijt}(\boldsymbol{\phi}_{it}^{(k)}) \\ \boldsymbol{\phi}_{it}^{(k)} &= x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)} \end{aligned}$$

then $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ *is a transfer minorization of* $\mathcal{L}(\boldsymbol{\theta})$.

Proof. See Appendix A.2. □

By Theorem 1, $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}), k = 0, 1, 2, ...$ converges to a local, if not global, maximum of $\mathcal{L}(\boldsymbol{\theta})$.

Equation 3 is a first-order Taylor expansion of the likelihood function. Note that the parameters that we search over in the iterative MM algorithm, i.e. $\boldsymbol{\theta}$ rather than $\boldsymbol{\theta}^{(k)}$, appear only in the third part of the right-hand side of Equation 3. Focusing on this third part, we can think of optimization of $S(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)})$ as a linear regression of $x_{it}\beta_j^{(k)}+\xi_{ijq(t)}^{(k)}+h_j\left(x_{it}\boldsymbol{\beta}^{(k)}+\boldsymbol{\xi}_{iq(t)}^{(k)};\boldsymbol{y}_{it}\right)$ on x_{jt} (with separate coefficients for each j) and $\xi_{ijq(t)}$. This is the central benefit of our MM approach to the multinomial logit. We convert non-linear optimization to sequential linear optimization, which is particularly attractive when there are many regressors. Rather than use OLS directly, we use linear panel data methods to address the large number of parameters represented by ξ_{ijq}.

Thus, the functional form of $h_j(\boldsymbol{\phi}_{it}^{(k)};\boldsymbol{y}_{it})$ is clearly important to our technique. For the multinomial logit, this function takes on a particularly simple form: $y_{ijt} - p_{ijt}(\boldsymbol{\phi}_{it}^{(k)})$. Thinking of $x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)}$ as the expectation of u_{ijt} at iteration k, our iterative linear regression uses a dependent variable above the expectation for observations with $y_{ijt} = 1$ and below the expectation for observations with $y_{ijt} = 0$.

Our full algorithm is as follows. We begin with a guess of the parameters $\boldsymbol{\theta}^{(1)}$. We iterate on the following sequence of steps, which updates the parameters $\boldsymbol{\theta}^{(k)}$ in iteration k:

1. Minorization step: Given $\boldsymbol{\theta}^{(k)}$, we calculate:
$$v_{ijt}^{(k)} = x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)} + h_j\left(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)};\boldsymbol{y}_{it}\right). \tag{4}$$
2. Maximization step:
   (a) Update $\boldsymbol{\beta}^{(k)}$ by demeaned OLS:
   $$\beta_j^{(k+1)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{x}_{it}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{x}_{it}'\tilde{v}_{ijt}^{(k)} \quad \forall j,$$
   where $\tilde{x}$ indicates demeaning at the consumer-quarter-product level.
   (b) Update $\{\boldsymbol{\xi}_{iq}^{(k)}\}_{i=1,...,N,q=1,...,Q}$ by computing:
   $$\xi_{ijq}^{(k+1)} = \frac{1}{T_q}\sum_{t\in\mathcal{T}_q}\left(v_{ijt}^{(k)} - x_{it}\beta_j^{(k+1)}\right) \quad \forall i, j, q,$$
   where T_q is the number of time periods in quarter q and $\mathcal{T}_q$ is the set of time periods in quarter q.
3. Return to Step 1 for iteration $k+1$ as long as the difference between $\boldsymbol{\theta}^{(k+1)}$ and $\boldsymbol{\theta}^{(k)}$ is above some tolerance.

4.3 Connection to the EM Algorithm for Binary Probit

If we could observe u_{ijt}, we could estimate our coefficients and fixed effects directly by linear techniques rather than relying on discrete outcome methods such as the multinomial logit. In this sense, Equation 4 in Step 1 of the MM algorithm above has the feel of *data augmentation* (Tanner,

1996). That is, we calculate v_{ijt} as an approximation of the unobserved u_{ijt}. That is the intuition behind many applications of the EM algorithm.

In the multinomial logit, v_{ijt} does not coincide with the expectation of u_ijt. That is, $v_{ijt}^{(k)} \neq E[u_{ijt}|x_{it}, y_{ijt}, \theta_j^{(k)}]$. Indeed, substituting $v_{ijt}^{(k)}$ with $E[u_{ijt}|x_{it}, y_{ijt}, \theta_j^{(k)}]$ would lead to biased results because the likelihood of the expectation is not equal to the expectation of the likelihood. Generating the expectation of the likelihood function for the case of the multinomial logit is more complicated than our minorization approach, so EM algorithm is relatively unattractive in our context.

However, the MM and EM algorithm coincide in the case of the binary probit. Indeed, Chen (2019) estimates the binary probit by the EM algorithm and derives a functional form equivalent to our MM algorithm. See also Greene (2018).

In the case of binary Probit,

$$l(x_{it}\beta + \xi_{iq(t)}; y_{it}) = \begin{cases} \Phi(x_{it}\beta + \xi_{iq(t)}) & y_{it} = 1 \\ 1 - \Phi(x_{it}\beta + \xi_{iq(t)}) & y_{it} = 0 \end{cases},$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\phi(\cdot)$ is its PDF.[5]

Accordingly,

$$\begin{aligned} h(x_{it}\beta + \xi_{iq(t)}; y_{it}) &= \frac{\partial \ln\left(x_{it}\beta + \xi_{iq(t)}; y_{it}\right)}{\partial x_{it}\beta + \xi_{iq(t)}} = \begin{cases} \dfrac{\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{\Phi\left(x_{it}\beta + \xi_i\right)}, & y_{it} = 1 \\ \dfrac{\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{1 - \Phi(x_{it}\beta + \xi_{iq(t)})}, & y_{it} = 0 \end{cases} \\ &= \frac{\left(y_{it} - \Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\right)\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{\Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\left(1 - \Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\right)} \end{aligned}$$

Thus, we see that $h(x_{it}\beta + \xi_{iq(t)}; y_{it})$ corresponds to the well-known Mills ratio and, as a result, $h(x_{it}\beta + \xi_{iq(t)}; y_{it}) = E[\epsilon_{it}|y_{it}]$, where ϵ_{it} corresponds to the normally distributed error from the probit model. In this sense, the MM and EM algorithms correspond in the case of the binary probit.

4.4 Simulation Results

Table 1: Comparison of $\hat{\beta}$ between MLE and MM-algorithm

	Binary probit	Binary logit	Multinomial logit	
	β	β	β_2	β_3
MLE	0.9697	1.0184	0.9983	0.4775
MM-algorithm	0.9697	1.0184	0.9983	0.4775

In this section, we show that our iterative algorithm gives the numerically identical result for $\hat{\beta}$ as standard MLE techniques. We generate data from three nonlinear models with a linear index: (1) a binary probit model, (2) a binary logit model and (3) a multinomial logit model. In each case, we estimate with the correct model by both traditional gradient-based techniques[6] and our MM algorithm and compare the result. We generate simulated data assuming that $N = 10$,

[5]Notation looks slightly different here because $J = 2$. It is no longer necessary to treat β, $\xi_{iq(t)}$ and y_{it} as vectors with only two choices.

[6]For the two binary choice models, we use the R function "glm" in package "stats", with its default method iterative reweighted least square (IWLS) ; for multinomial logit, we use the R function "mlogit" in package "mlogit", with its default method Broyden–Fletcher–Goldfarb–Shanno (BFGS)

$T = 1000$, $J = 2$ for binary choice models and $J = 3$ for the multinomial logit. We assume there are consumer-level fixed effects in all models, generated independently from the standard normal distribution. We do not add a time element to the fixed effects in this exercise. We let x_{it} be a scalar, also drawn independently from the standard normal distribution. For the two binary choice models, we set the true parameter $\beta = 1$; for the multinomial logit model, we normalize $\beta_1 = 0$ and $\xi_{i1} = 0, \ \forall i$, and set the true parameters $\beta_2 = 1$ and $\beta_3 = 0.5$. The results are summarized in Table 1.

5 Data

For data, we rely on the Nielsen Consumer Panel Dataset available through the Kilts Center for Marketing at the Chicago Booth School of Business. The Nielsen Consumer Panel provides detailed coverage of consumer purchase choices. In addition to scanning UPC symbols from products they bring home, consumers indicate how they paid for a product and submit their receipts from which Nielsen attempts to verify their choices.

We study three choices: cash, card and check. For some issues, it may be interesting to distinguish between credit card and debit card. However, debit cards may be authorized by signature or PIN (a personal 4 to 6-digit number) and many consumers do not understand the difference between signature debit and credit cards (which are authorized only by signature in the United States).[7] As a result, we treat credit and debit as a single choice: card. In general, debit cards and credit cards have similar efficiency levels as they are both electronic

Although the Nielsen panel runs over a decade, Kilts gained access to the payments data only since 2013. Thus, we have five years of data. In those five years, we observe 77,657 households and 31,344,415 trips. That is 403.6 trips per household on average. There is some turnover in households in the data set, so the average number of years that we observe a household is 2.53 years. We observe that 19.9% of households in the data set stayed for the entire five years, with 31.0% staying longer than four years and 44.3% staying longer than three years[8].

Overall, we observe 153.7 trips per household per year on average.[9] We observe 36.4 trips per household quarter, with a 25th percentile of 18 and a 75th percentile of 50. As our most granular specification uses two fixed effects per household-quarter in a non-linear model, the incidental parameters problem is potentially an issue with these numbers of observations.

Over our entire data set, the market share for transactions for cash, check and card is 30.9%, 2.4% and 66.8%.[10] These shares vary substantially over time and across different types of consumers and transactions. Figure 2 shows how transactions by payment type vary over time. We see that there has been a substantial increase in card use over time, and an increase in the total number of transactions as well.

Table 2: Transaction size distribution ($)

Mean	Std. Err.	10%	[25%, 75%]	90%
46.84	65.10	5.22	[11.94, 57.06]	107.51

We focus particularly on the role of transaction size. In Table 2, we see that the average

[7] Past versions of the Nielsen panel appeared to give consumers contradictory instructions on this issue, for instance, instructing consumers to indicate "credit" if they used a signature. We found it difficult to verify the instructions for the current data set. See Cohen, Rysman & Wozniak (2017).

[8] To be consistent with the model with quarterly household-method fixed effects, we first counted how many quarters that each household stayed, and then converted it into years by dividing it by four.

[9] Among these trips, 34.6% involve grocery stores, other top types of retailers include discount stores (15.7%), drug store(6.0%), dollar store (4.9%), warehouse club (4.9%), Quick Serve Restaurants (3.1%), etc.

[10] Two types of payment methods in the raw data are excluded from our estimation, they are "Scanner does not collect Method of Payment" and "Other Payment" which accounts for 40.3% and 1.71% of the trips in the raw data.

transaction size in our data is $46.84. The variation in large, with the 10th percentile at $5.22 and the 90th at $107.51. The interquartile range is [$11.94, $57.06].

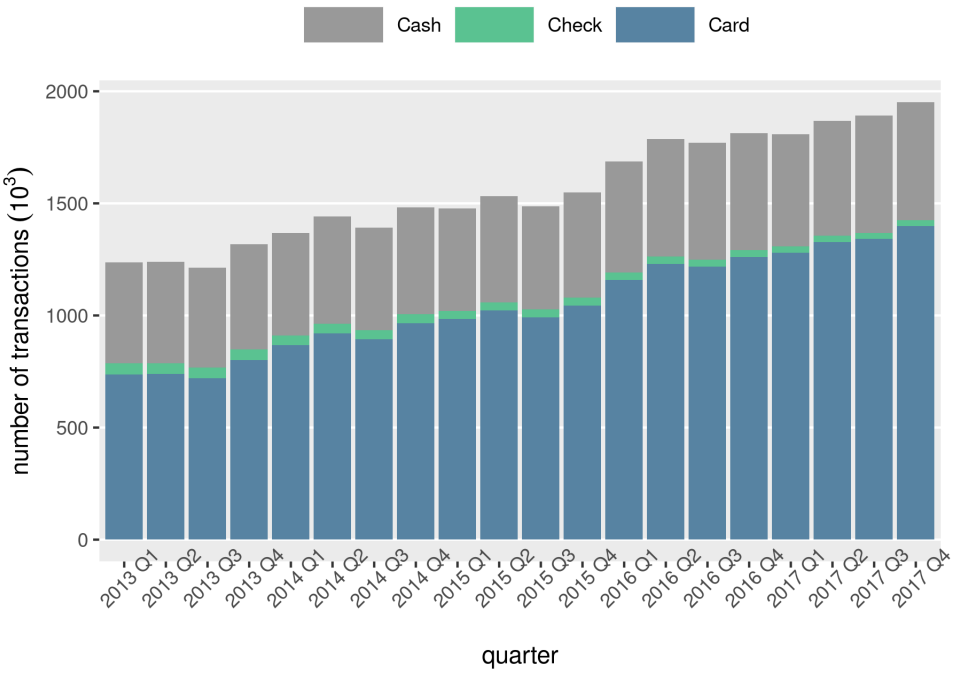

Figure 2: Transactions over time

Transaction size is an important determinant of payment choice. Figure 3 shows a dramatic change in market shares as transaction size changes, with the share of cash transactions moving from above 60% to below 20% as transaction size moves from $5 to $150. Most of the remaining share is absorbed by card, although check also increases in share as transaction size increases, obtaining close to 4% of transactions for the largest transaction sizes. Although check has a low market share, 37.6% of the households in our data used check for at least once, so it is prevalent in household decision-making.

6 Results

We first present our model in the context of payments. Consumers exogenously need to shop for a basket of goods for a given transaction size, for which the consumer must choose a payment mechanism. In particular, household i on shopping trip t that takes place in quarter $q(i,t)$ that costs x_{it} dollars paid with payment mechanism $j \in \{\texttt{cash}, \texttt{check}, \texttt{card}\}$ receives utility:

$$u_{ijt} = \beta_j \ln(x_{it}) + \xi_{ijq(i,t)} + \varepsilon_{ijt}.$$

Relative to Equation 1, we take the log of x_{it} and emphasize that x_{it} is a scalar. Note that in our approach, t indexes shopping trips rather than calendar time, so two households may be in different quarters for the same shopping trip number. Thus, rather than write $q(t)$ as in Equation 1, we write $q(i,t)$. As above, ε_{ijt} is distributed Extreme Value.

As is standard, we must normalize the mean utility of one choice to zero. We normalize the utility of $j = \texttt{cash}$ to zero, so $\beta_{\texttt{cash}} = 0$ and $\xi_{i,\texttt{cash},q(i,t)} = 0$ for all i and q. We interpret the rest of the coefficients as the value relative to the value for cash.

With previous research in mind, we are particularly interested in comparing our model to one where we are restricted not to be able to track consumers. That is, we compare our results to

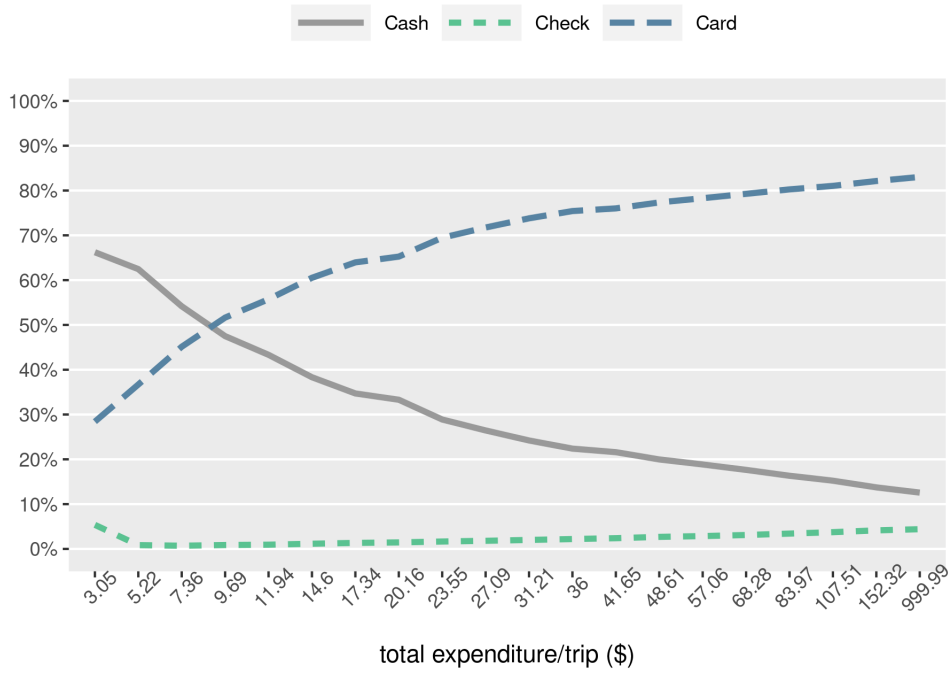

Figure 3: Market share in transactions by transaction size

the case where we assume $\xi_{ijq(i,t)} = \xi_j$ for all i and q. For comparison purposes, we also consider household-product fixed effects, so $\xi_{ijq} = \xi_{ij}$ for all q.

Results appear in Table 3. Standard errors in this table are conventional maximum likelihood standard errors derived from the inverse of the Hessian matrix. As the Hessian is very large, we exploit the sparsity of the matrix in order to invert it, as presented in Appendix B.

Before discussing parameters, it is worth considering how long it takes to estimate these models. For the case of only product fixed effects (the first panel), we estimate the likelihood model by both our MM algorithm and by a traditional gradient optimization routine, BFGS. Results are numerically very close, especially for $\beta_{\texttt{card}}$, most likely because we have much higher market share for card than check. Our experiments show the remaining differences can be reduced by reducing the tolerance levels in the optimizers. In the last row, we see that the MM algorithm estimates the model in about half the time. For the case with household-product fixed effects, the MM algorithm takes about 13 hours and with household-quarter-product fixed effects, it takes about 24 hours. In contrast, we ran the BFGS algorithm using a dummy variable approach to implementing the fixed effects for these two cases and never reached convergence for either case.[11] We do not report BFGS results for these cases.

	ξ_j			ξ_{ij}			ξ_{iqj}	
	MM	BFGS		MM	BFGS		MM	BFGS
β_{check}	0.7188	0.7278	β_{check}	1.0730		β_{check}	1.1595	
	(0.0012)	(0.0012)		(0.0017)			(0.0029)	
β_{card}	0.7085	0.7085	β_{card}	1.0563		β_{card}	1.1570	
	(0.0004)	(0.0004)		(0.0006)			(0.0011)	
ξ_{check}	-4.8441	-4.8760	$\bar{\xi}_{check}$	-7.7561		$\bar{\xi}_{check}$	-7.9493	
			$[min, max]$	[-15.2902, 6.5288]		$[min, max]$	[-16.0147, 7.4957]	
ξ_{card}	-1.4628	-1.4627	$\bar{\xi}_{card}$	-2.1049		$\bar{\xi}_{card}$	-2.0048	
			$[min, max]$	[-15.1744, 6.6956]		$[min, max]$	[-15.9977, 8.0039]	
Number of FE's estimated	N = 2			N = 155,314			N = 1,631,808	
time	$\sim$ 30 min	$\sim$ 1 hr		$\sim$ 13 hr	$>$ 24 hr		$\sim$ 24 hr	$>$ 24 hr

Notes: All estimations were run on a CPU with 8 8G-memory processors

Table 3: Estimates & Computational Time: MM vs. BFGS

[11] We write "greater than 24 hours" in the table, but in practice, we let the program run for at least twice this.

Comparing across columns, we see that the coefficients grow on transaction size for check and card as we add household and household-quarter fixed effects, with a bigger change in going from product to product-household fixed effects than from going from product-household to product-household-quarter fixed effects. However, at the same time, the fixed effects change, so it is difficult to say whether how the marginal effect of transaction size changes with the specification. For this, we turn to computing average marginal effects.

Given our multinomial logit assumption, we define the average marginal effect (AME) of x_{it} on the probability of choice $p_{ijt}(\theta)$ as:

$$\text{AME}_{ijt} = \frac{1}{NT}\sum_{i=1,t=1}^{N,T}\frac{\partial p_{ijt}(\theta)}{\partial x_{it}} = \frac{1}{NT}\sum_{i=1,t=1}^{N,T}\frac{p_{ijt}(\theta)\left(\beta_j - \sum_{k=1}^{3}\beta_k p_{ikt}(\theta)\right)}{x_{it}}.$$

	ξ_m			ξ_{mh}			ξ_{mhq}		
	cash	check	card	cash	check	card	cash	check	card
$\hat{AME}$	-0.0107	0.0004	0.0103	-0.0101	0.0005	0.0096	-0.0100	0.0005	0.0095
$\tilde{AME}$				-0.0099	0.0005	0.0094	-0.0097	0.0005	0.0092
Bias: $\hat{AME}$ - $\tilde{AME}$				-0.0002	0.0000	0.0002	-0.0003	0.0000	0.0003

Table 4: Average marginal effects of transaction size

Results appear in Table 4. The table shows that the specifications with product-household and product-household-quarter fixed effects (panels 2 and 3) lead to very similar AMEs. However, there is a large change when comparing these results to the first panel, the specification with product fixed effects. The AME is about half as large in the cases with more fixed effects. Intuitively, some households have large transaction sizes and almost always use check or card. The first specification interprets this cross-sectional heterogeneity to say that transaction size has a relatively large effect on payment choice. However, when we focus on within-household heterogeneity by including product-household (or product-household-quarter) fixed effects, the correlation between transaction size and payment choice is smaller. In these specifications, doubling the transaction size reduces the probability of using cash by about 1%, with most of that going to card.

The parameters we have presented so far are subject to bias due to the incidental parameters problem. To address this, we implement the split-panel jackknife of Dhaene & Jochmans (2015). They recommend partitioning the data set into two halves based on time. In our setting, we split each household's trips into two halves. That means that each half of the data set has households for different numbers of observations and a given week may appear in both halves of the data set. We re-estimate the model on each half of the data set. Let $\hat{\boldsymbol{\beta}}$ denote the estimates from the full sample, and $\hat{\boldsymbol{\beta}}^1$ and $\hat{\boldsymbol{\beta}}^2$ be the estimates from each half, so:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^1 &= \arg\max \textstyle\sum_{i,t\in[1,\lfloor T_i/2\rfloor]} \ln l\left(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq}(\boldsymbol{\beta}); \boldsymbol{y}_{iqt}\right)\\ \hat{\boldsymbol{\beta}}^2 &= \arg\max \textstyle\sum_{i,t\in[\lfloor T_i/2\rfloor+1,T_i]} \ln l\left(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq}(\boldsymbol{\beta}); \boldsymbol{y}_{it}\right)\end{aligned}$$

Here, we write $\boldsymbol{\xi}_{iq(i,t)}(\boldsymbol{\beta})$ to emphasize that the estimates of the fixed effects will change in both estimations. With these results, the bias-corrected estimates are:

$$\tilde{\boldsymbol{\beta}} = 2\hat{\boldsymbol{\beta}} - \frac{\hat{\boldsymbol{\beta}}^1 + \hat{\boldsymbol{\beta}}^2}{2}.$$

The results appear in Table 5. The bias reduction reduces the parameter estimate of both $\beta_{\texttt{check}}$ and $\beta_{\texttt{card}}$, but by a small amount, less than 4% in each case. As we stated above, the average number of trips per quarter is 36.4. Evidently, this is enough to significantly mitigate the incidental parameters problem, even with three choices (that is, two fixed-effects per household-quarter).

	ξ_{ij}		ξ_{iqj}	
	β_{check}	β_{card}	β_{check}	β_{card}
$\hat{\beta}$	1.0730	1.0563	1.1595	1.1570
$\tilde{\beta}$	1.0519	1.0319	1.1247	1.1146
Bias: $\hat{\beta} - \tilde{\beta}$	0.0211	0.0244	0.0348	0.0425

Table 5: Split-panel Jackknife Correction

7 Long-term decomposition

We can see in Figure 2 that the share of card usage is growing over time. This growth in share could be due either to a change in choice probabilities within households or a change in the composition of transactions or transaction sizes across households. To fix intuition, suppose there is a young household that always uses a card for payments and an older household that always uses cash or check. If the young household has children, its number of shopping trips and transaction sizes will tend to grow. If the older household reaches retirement age, its number of shopping trips and transaction sizes will tend to shrink. Even without any changes in preferences, these compositional changes will lead to an increase in the share of card use as measured by transactions or transaction values. Naturally, if the older household leaves the sample and is replaced by a younger, more heavily card-using household, that will further contribute to compositional change.

The goal of this section is to decompose the change in card share into changes in choice probability within households and changes in the composition of payments across households. To facilitate discussion, we introduce some new notations. We denote $\mathcal{I}_q$ as the set of households in quarter q, and $\mathcal{T}_{iq}$ as the set of trips of household i in quarter q. The transactions market share of payment choice j in quarter q is:

$$s_{jq} = \frac{1}{\sum_{i \in \mathcal{I}_q} |\mathcal{T}_{iq}|} \sum_{i \in \mathcal{I}_q, t \in \mathcal{T}_{iq}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijq(i,t)}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikq(i,t)}\right)}.$$

The market share s_{jq} can change over time for a number of reasons: the number of transactions $|\mathcal{T}_{iq}|$ can change, the values of those transactions x_{it} can change, preferences $\xi_{ijq(i,t)}$ can change, or the set of households $\mathcal{I}_q$ can change, which can be further broken down into entry and exit. We proceed by sequentially fixing each of these values at their realization in the first quarter that each household i is observed in the data, denoted as $\underline{q}(i)$, or in the case of exit the last quarter, denoted as $\overline{q}(i)$, and then computing market shares for the last quarter.

Transaction size distribution within households: For each household, we fix the number of trips and the total expenditure on each trip at the level of their first quarter, but let their fixed effects evolve with time. We calculate the household-level choice probabilities and then aggregate them to market shares with the number of trips in the current quarter as weights. So the counterfactual last quarter market share is:

$$s^1_{jQ} = \frac{1}{\sum_{i \in \mathcal{I}_Q} |\mathcal{T}_{iQ}|} \sum_{i \in \mathcal{I}_Q} \frac{|\mathcal{T}_{iQ}|}{|\mathcal{T}_{i\underline{q}(i)}|} \sum_{t \in \mathcal{T}_{i\underline{q}(i)}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikQ}\right)}. \tag{5}$$

Consider the case in which the set of transactions sizes realized in $\underline{q}(i)$ was the same as in Q. That would imply that the number of transactions in each period were the same, so $|\mathcal{T}_{iQ}| = |\mathcal{T}_{i\underline{q}(i)}|$, and the set of x_{it} was the same for the first and last period that i was in the data. In this case, $s_{jQ} = s^1_{jQ}$. The difference $s_{jQ} - s^1_{jQ}$ provides a measure of how changes in the distribution of transactions contributes to the change in market share $s_{jQ} - s_{j1}$.

Product-household-quarter fixed effects: We capture the change in preferences within households with our product-household-quarter fixed effects. In order to mute the effect of preferences, we fix product-household-quarter fixed effects at the level of the first quarter the household is observed and then calculate the market share in the final quarter Q as:

$$s^2_{jQ} = \frac{1}{\sum_{i\in\mathcal{I}_Q}|\mathcal{T}_{iQ}|}\sum_{i\in\mathcal{I}_Q, t\in\mathcal{T}_{iQ}}\frac{\exp\left(x_{it}\beta_j + \xi_{ij\underline{q}(i)}\right)}{\sum_{k=1}^{J}\exp\left(x_{it}\beta_k + \xi_{ik\underline{q}(i)}\right)}. \tag{6}$$

In this case, $s_{jQ} - s^2_{jQ}$ provides a measure of the contribution of changes in product-household-choice fixed effects, and this term equals zero only if fixed effects are the same in the first and last perod.

Number of transactions across households: As in the young household vs. older household example discussed at the beginning of this section, the growth of card usage could also be due to shifts in transactions from non-card to card users. To isolate this effect, we first calculate the household-level choice probabilities, and when aggregating them to compute market share, we weight by the number of trips in the household's first quarter rather than the number of trips in the current quarter. Then, the last-quarter market share becomes:

$$s^3_{jQ} = \frac{1}{\sum_{i\in\mathcal{I}_Q}|\mathcal{T}_{i\underline{q}(i)}|}\sum_{i\in\mathcal{I}_Q}\frac{|\mathcal{T}_{i\underline{q}(i)}|}{|\mathcal{T}_{iQ}|}\sum_{t\in\mathcal{T}_{iQ}}\frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J}\exp\left(x_{it}\beta_k + \xi_{ikQ}\right)}. \tag{7}$$

Entry: In this scenario, we focus on those households that stay in the data from the first quarter. They are allowed to exit as observed in the data, in order to be distinguished from the exit channel. Specifically, starting with the 31,178 households in 2013 Q1, 15,477 of them stay until the last quarter 2017 Q4, which is 32.6% of all households at that time. The market share in the final quarter when fixed $\mathcal{I}_Q = \mathcal{I}_1$:

$$s^4_{jQ} = \frac{1}{\sum_{i\in\mathcal{I}_1}|\mathcal{T}_{iQ}|}\sum_{i\in\mathcal{I}_1, t\in\mathcal{T}_{iQ}}\frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J}\exp\left(x_{it}\beta_k + \xi_{ikQ}\right)} \tag{8}$$

Exit: We consider a counterfactual scenario where no households leave the sample. Therefore, all households that ever show up in the sample stay until the last quarter 2017 Q4, i.e. $\mathcal{I}_Q = \bigcup_{q=1}^{Q}\mathcal{I}_q$, which gives 77,656 households. For those households that leave before 2017 Q4, we assume that their number of trips, transaction size of each trip and fixed effects are the same as in the last quarter that they are observed in the data, i.e. $\mathcal{T}_{iq} = \mathcal{T}_{i\overline{q}(i)}, \forall q > \overline{q}(i)$. In particular, $\overline{q}(i) = Q$ for household i that is observed in 2017 Q4. In this scenario, the market share in the final quarter is:

$$s^5_{jQ} = \frac{1}{\sum_{i\in\bigcup_{q=1}^{Q}\mathcal{I}_q}|\mathcal{T}_{i\overline{q}(i)}|}\sum_{i\in\bigcup_{q=1}^{Q}\mathcal{I}_q, t\in\mathcal{T}_{i\overline{q}(i)}}\frac{\exp\left(x_{it}\beta_j + \xi_{ij\overline{q}(i)}\right)}{\sum_{k=1}^{J}\exp\left(x_{it}\beta_k + \xi_{ik\overline{q}(i)}\right)} \tag{9}$$

Then, the contribution of each channel is the difference $s_{jQ} - s^k_{jQ}$ for each $k = \{1, 2, 3, 4, 5\}$. Note that the sum of these differences does not exactly equal $s_{jQ} - s_{j1}$, in part because of joint effects. By isolating each effect separately, we do not capture the role of simultaneous changes in channels, for instance because in practice, ξ_{ijq} and x_{it} change jointly. Still, these differences give a first-order approximation of relatively how much each type of change contributes to the overall change. Therefore, for demonstration purpose, we rescale these differences so that the sum of them

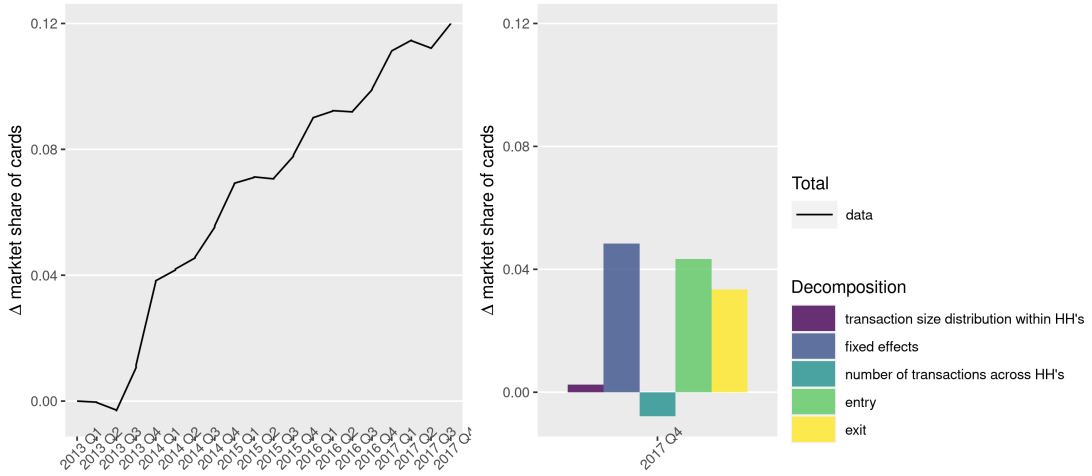

Figure 4: Long-term decomposition

equals to $s_{jQ} - s_{j1}$.[12]

The result appears in Figure 4. We see that changes in preferences are the single biggest driver of the change in market share. However, that still accounts for only about one third the change. Entry and exit are other large contributors. In fact, changes in the number of transactions contribute *against* the overall change. That is, we find that the number of transactions has distributed slightly against card-users. Overall, it appears that changes in preferences are a major contributor to changes in market share, but hardly the only one.

8 Conclusion

The transition to digital forms have payments has been one of the most visible and important examples of digitization in the economy and one with significant policy implications. We utilize a new source of data on payment behavior: consumer panel surveys. Although these sources are more associated with studying consumer shopping behavior and responses to advertising, we show that these data can be usefully employed to study payment behavior. Unlike previous studies, we observe consumer payment choice at the level of the trip at high frequency over a long period for many households.

Thus, we can track household behavior over time and account for household heterogeneity using panel data techniques. While multinomial models of choice are natural in payment studies, they are challenging in this context because non-linear estimation can become infeasible in the face of millions of fixed effects. In our preferred specification, we have more than 11 million fixed effects. We present a new method for addressing fixed effects in multinomial models based on the Minorization-Maximization (MM) algorithm, which can be seen as a generalization of the EM algorithm. While the MM algorithm has a significant history in the statistics literature, we are aware of almost no presence in econometrics. We discuss the application of our method not only to the multinomial logit model, but also several other well-known models.

Utilizing our method, we are able to find maximum likelihood estimates of our multinomial choice problem. We find that while transaction size is an important determinant of payment choice, its effect is overstated in previous research that did not utilize household fixed effects. We apply our results to study the increase in card usage in our data, decomposing the change into that due to changes in household preferences (in our case, changes in household-quarter-choice fixed effects)

[12] In this sense, our measure is similar to Variance Partition Coefficients, as in Goldstein, Browne & Rasbash (2002). See also Grömping (2007).

and changes in the composition of payments across households, that is, changes in the number of transactions and transactions sizes across households. We find that changes in preferences account for only a third of the observed change in card usage. Thus, the primary driver in changes in payment usage are the decline in transactions and transaction sizes for non-card using households.

References

Arango, C., Huynh, K., & Sabetti, L. (2015). Consumer payment choice: Merchant card acceptance versus pricing incentives. *Journal of Banking and Finance*, *55*, 130–141.

Böhning, D. & Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statical Mathematics*, *40*, 641–663.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, *47*, 225–238.

Chen, M. (2019). Estimation of nonlinear panel models with multiple unobserved effects. Unpublished manuscript, University of Warwick.

Cohen, M., Rysman, M., & Wozniak, K. (2017). Payment choice with consumer panel data. Unpublished Manuscript.

de Leeuw, J. & Lange, K. (2009). Sharp quadratic majorization in one dimension. *Computational Statistics & Data Analysis*, *53*, 2471–2484.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statisical Society, Series B*, *39*, 1–38.

Dhaene, G. & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies*, *82*, 991–1030.

Dutkowsky, D. & Fusaro, M. (2011). What explains consumption in the very short run? Evidence from checking account data. *Journal of Macroeconomics*, *33*, 542–552.

D'Haultfœuille, X. & Iaria, A. (2016). A convenient method for the estimation of the multinomial logit model with fixed effects. *Economics Letters*, *141*, 77–70.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences*, *1*(4), 223–231.

Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Prentice Hall.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, *61*, 139–147.

Hahn, J. & Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, *72*, 1295–1319.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Hinz, J., Hudle, A., & Wanner, J. (2019). Separating the wheat from the chaff: Fast estimation of GLMs with high-dimensional fixed effects. Unpublished manuscript, European University Institute.

Ho, K. & Pakes, A. (2014). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, *104*, 3841–3884.

Hospido, L. (2012). Modelling heterogeneity and dynamics in the volatility of individual wages. *Journal of Applied Econometrics*, *27*, 386–414.

Hunter, D. R. & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *58*, 30–37.

James, J. (2017). MM algorithm for general mixed multinomial logit models. *Journal of Applied Econometrics*, *32*, 841–857.

Jonker, N. & Kosse, A. (2009). The impact of survey design on research outcomes: A case study of seven pilots measuring cash usage in the Netherlands. Working Paper 221/2009 Bank of Netherlands.

Klee, E. (2008). How people pay: Evidence from grocery store data. *Journal of Monetary Economics*, *55*, 526–541.

Koulayev, S., Rysman, M., Schuh, S., & Stavins, J. (2016). Explaining adoption and use of payment instruments by US consumers. *RAND Journal of Economics*, *47*, 293–325.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM.

Lange, K., Hunter, D. R., & Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, *9*, 1–20.

McFadden, D. & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

Rysman, M. (2007). Empirical analysis of payment card usage. *Journal of Industrial Economics*, *60*, 1–36.

Schuh, S. & Stavins, J. (2010). Why are (some) consumers (finally) writing fewer checks? The role of payment characteristics. *Journal of Banking and Finance*, *34*, 1745 – 1758.

Shum, M., Song, W., & Shi, X. (2018). Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. *Econometrica*, *86*, 737–761.

Stamann, A. (2018). Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects. Unpublished manuscript, arXiv:1707.01815.

Stammann, A., Heiß, F., & McFadden, D. (2016). Estimating fixed effects logit models with large panel data. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V3.

Stango, V. & Zinman, J. (2014). Limited and varying consumer attention: Evidence from shocks to the salience of bank overdraft fees. *Review of Financial Studies*, *27*, 990–1030.

Tanner, M. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer.

Wakamori, N. & Welte, A. (2017). Why do shoppers use cash? Evidence from shopping diary data. *Journal of Money, Credit and Banking*, *2017*, 115–169.

Wang, Z. & Wolman, A. F. (2016). Payment choice and currency use: Insights from two billion retail transactions. *Journal of Monetary Economics*, *84*, 94–115.

White, K. J. (1975). Consumer choice and use of bank credit cards: A model and cross-section results. *Journal of Consumer Research*, *2*, 10–18.

Appendices

Appendix A Proofs

A.1 Proof of Theorem 1

(1) Given that $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ and by definition of minorization for maximization,

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) \geq S(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq S(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{10}$$

Because $\mathcal{L}(\boldsymbol{\theta})$ is a log likelihood function, we have that $\mathcal{L}(\boldsymbol{\theta}) \leq 0, \forall \boldsymbol{\theta}$. Then the fact that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ is an increasing sequence bounded above implies its convergence to some $\mathcal{L}^* \leq 0$. Hence for any $\delta > 0$, there exists a $p(\delta)$ such that for all $p \geq p(\delta)$ and all $r \geq 1$,

$$\sum_{s=1}^{r}\{\mathcal{L}(\boldsymbol{\theta}^{(p+s)}) - \mathcal{L}(\boldsymbol{\theta}^{(p+s-1)})\} = \mathcal{L}(\boldsymbol{\theta}^{(p+r)}) - \mathcal{L}(\boldsymbol{\theta}^{(p)}) < \delta \tag{11}$$

From Eq.(10), we have

$$S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(p+s)}) - \mathcal{L}(\boldsymbol{\theta}^{(p+s-1)}),\ \forall s \geq 1, \tag{12}$$

and by Taylor expansion,

$$\begin{aligned} & S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) \\ = \ & -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{10}S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) \\ & \quad -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}) \\ = \ & -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}) \end{aligned} \tag{13}$$

where $\boldsymbol{\theta}_0^{(p+s)}$ is some point on the line segment joining $\boldsymbol{\theta}^{(p+s-1)}$ and $\boldsymbol{\theta}^{(p+s)}$. The second equality holds because $\nabla^{10}S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) = 0$ is the necessary condition for $\boldsymbol{\theta}^{(p+s)} = \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta};\boldsymbol{\theta}^{(p+s-1)})$ given that $S(\boldsymbol{\theta};\boldsymbol{\theta}')$ is twice differentiable by definition.

Furthermore, Definition 1 (2) indicates that $\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})$ is negative definite, i.e.. $-\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})$ is positive definite. Therefore, let I be an identity matrix, $\exists\lambda > 0$ such that $-\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) > \lambda I$, i.e. $\nabla^{20}S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) + \lambda I$ is positive definite.

Then Eq.(13) can be rewritten as

$$S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) > \lambda(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}), \forall s \geq 1 \tag{14}$$

Combining Eq.(11) , Eq.(12), Eq.(13) and Eq.(14), we have

$$\begin{aligned} & \lambda\sum_{s=1}^{r}(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}) \\ < \ & \sum_{s=1}^{r} S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) \\ < \ & \delta \end{aligned} \tag{15}$$

for all $p \geq p(\delta)$ and all $r \geq 1$, which proves $\boldsymbol{\theta}^{(k)}$ converges to some $\boldsymbol{\theta}^*$ in the closure of Ω.

(2) Since $\boldsymbol{\theta}^{(k)}, k = 0, 1, 2, ...$ converges to $\boldsymbol{\theta}^*$,

$$\boldsymbol{\theta}^* = \arg\max S(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$$

Then by Corollary 1,

$$\mathcal{L}(\boldsymbol{\theta}^*) = \nabla^{10}S(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*) = 0$$

Similarly, $\nabla^{20}S(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$ is negative definite. Q.E.D

A.2 Proof of Theorem 2

We show that $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ is a minorization of $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ at $\boldsymbol{\theta}^{(k)}$ for maximization by checking the three requirements in Definition 1.

(1) $\underline{S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(k)})}$

We start from a representative consumer i. Recall that $\boldsymbol{\phi}_{it} = x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}$. Then, in the case of multinomial logit,

$$l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) = \prod_j \Big(\frac{exp(\phi_{ijt})}{1+\sum_{k\neq j} exp(\phi_{ikt})}\Big)^{y_{ijt}}$$

$$\begin{aligned} h_j(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) &= \frac{\partial \log\ l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it})}{\partial \phi_{ijt}} \\ &= y_{ijt} - p_{ijt} \end{aligned}$$

$$\begin{aligned} h_{jk}(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) &= \frac{\partial^2 \log\ l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it})}{\partial \phi_{ijt}\partial \phi_{ikt}} \\ &= \begin{cases} -p_{ijk}(1-p_{ikt}), & j=k \\ p_{ijt}p_{ikt}, & j\neq k \end{cases} \end{aligned}$$

The Taylor expansion of $log\ l(\boldsymbol{\phi}_{it})$ at $\tilde{\boldsymbol{\phi}}_{it}$ is ,

$$\log\ l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) = \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + (\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it})'\nabla \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + \frac{1}{2}(\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it})'\nabla^2 \log\ l(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it})(\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it}) \tag{16}$$

Given that

$$\begin{aligned} \nabla^2 \log\ l(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) &= \begin{bmatrix} h_{11}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) & h_{12}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) & ... & h_{1J}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) \\ h_{21}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) & h_{22}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) & ... & h_{2J}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) \\ \vdots & \vdots & & \vdots \\ h_{J1}(\boldsymbol{\phi}^*_{it}) & h_{J2}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) & ... & h_{JJ}(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) \end{bmatrix} \\ &= \begin{bmatrix} -p_1(1-p_1) & p_1p_2 & ... & p_1p_J \\ p_2p_1 & -p_2(1-p_2) & ... & p_2p_J \\ \vdots & \vdots & & \vdots \\ p_Jp_1 & p_1p_2 & ... & -p_J(1-p_J) \end{bmatrix}, \end{aligned} \tag{17}$$

we have $\nabla^2 \log\ l(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) \geq -I$, i.e. $\nabla^2 \log\ l(\boldsymbol{\phi}^*_{it};\boldsymbol{y}_{it}) + I$ is semi-positive definite matrix.

Therefore,

$$\begin{aligned} \log\ l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) &\geq \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + (\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it})'\nabla \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) - \frac{1}{2}(\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it})'(\boldsymbol{\phi}_{it}-\tilde{\boldsymbol{\phi}}_{it}) \\ &= \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + \sum_j h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it})(\phi_{ijt}-\tilde{\phi}_{ijt}) - \frac{1}{2}\sum_j(\phi_{ijt}-\tilde{\phi}_{ijt})^2 \\ &= \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + \frac{1}{2}\sum_j h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it})^2 - \frac{1}{2}\sum_j(\tilde{\phi}_{ijt} - h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) - \phi_{ijt})^2 \end{aligned} \tag{18}$$

Recall that $l(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it})$ is the individual log-likelihood, the log-likelihood function is its sum over i and t. and we substitute $\boldsymbol{\phi}_{it}$ and $\tilde{\boldsymbol{\phi}}_{it}$ in inequality (18) with $x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}$ and $x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}$ respectively. It gives

$$\mathcal{L}(\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}; \boldsymbol{y}_{it})^2 - \frac{1}{2}\sum_{i,j,t}(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}; \boldsymbol{y}_{it}) - x_{it}\boldsymbol{\beta} - \boldsymbol{\xi}_{iq(t)})^2 \tag{19}$$

where the RHS of the above inequality as $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$.

(2) $\underline{S(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)})}$

By definition,

$$\begin{aligned} S(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}; \boldsymbol{y}_{it})^2 - \\ &\qquad \frac{1}{2}\sum_{i,j,t}(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}; \boldsymbol{y}_{it}) - x_{it}\boldsymbol{\beta}^{(k)} - \boldsymbol{\xi}^{(k)}_{iq(t)})^2 \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}; \boldsymbol{y}_{it})^2 - \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)}; \boldsymbol{y}_{it})^2 \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) \end{aligned}$$

(3) $\underline{\nabla^{20} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \text{ exists, and } \nabla^{20} S(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \text{ is negative definite}}$

To ease notation, we consider $\boldsymbol{\xi}_{iq(t)}$ as coefficients on indicator variables, combine these indicator variables with x_{it} and denote the combined vector as z_{ijt}. Then by the definition of $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ in Eq.(19),

$$\nabla^{20} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = -2\boldsymbol{Z}'\boldsymbol{Z}, \ \forall \boldsymbol{\theta}$$

where $\boldsymbol{Z}$ is the matrix with $z_{ijt}, \ \forall i, j, t$ stacked by rows.

Q.E.D

Appendix B Computing Standard Errors

We rewrite the likelihood function as $l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(i,t)}; \boldsymbol{y}_{it}) = l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})$. For each household i and each trip t,

$$\begin{aligned} \frac{\partial \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\theta}} &= \frac{\partial \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\phi}_{it}} \times \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}} \\ &= (\boldsymbol{y}_{it} - \boldsymbol{p}_{it})' \times \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}} \end{aligned} \tag{20}$$

where $\dfrac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}}$ is a $J \times (J + J \times I \times Q)$ matrix.

Let $\text{diag}_J(a)$ be a $J \times J$ matrix with diagonal elements as a and off-diagonal elements as 0,

$$\begin{array}{lcll}\left[\frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}}\right]_{.;1:J} & = & \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\beta}} & = \text{diag}_J(x_{it}) \\ \left[\frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}}\right]_{.;J\times i+1:J} & = & \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\xi}_{iq(i,t)}} & = \text{diag}_J(1)\end{array}$$

while the other elements are zero because $\boldsymbol{\xi}_{hq(h,t)}$ is not in $l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})$ if $h \neq i$.

Therefore, Equation 20 can be rewritten as:

$$((y_{i1t} - p_{i1t})x_{it}, \cdots, (y_{iJt} - p_{iJt})x_{it}, 0, \cdots, 0, y_{i1t} - p_{i1t}, \cdots, y_{iJt} - p_{iJt}, 0, \cdots, 0)$$

Then $\frac{\partial^2 \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ matrix can be written as

$$\begin{bmatrix} H_{\boldsymbol{\phi}_{it}} x_{it}^2, & \mathbf{0}, & \cdots, & \mathbf{0} & H_{\boldsymbol{\phi}_{it}} x_{it} & \mathbf{0}, & \cdots, & \mathbf{0} \\ \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\ H_{\boldsymbol{\phi}_{it}} x_{it} & \mathbf{0}, & \cdots, & \mathbf{0} & H_{\boldsymbol{\phi}_{it}} & \mathbf{0}, & \cdots, & \mathbf{0} \\ \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \end{bmatrix}$$

where $H_{\boldsymbol{\phi}_{it}}$ is as defined in Eq.(17)

And then the Hessian matrix is $\frac{1}{I \times T} \sum_{i,t} \frac{\partial^2 \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$