Institute for
New Economic Thinking
AT THE OXFORD MARTIN SCHOOL

Estimating Very Large Demand Systems

Joshua Lanier, Jeremy Large and John Quah

27th June 2022

INET Oxford Working Paper No. 2023-01

Estimating Very Large Demand Systems

Joshua Lanier* Jeremy Large† John Quah‡

27 June 2022

The latest version of this paper is available here

Abstract

We present a discrete choice, random utility model and a new estimation technique for analyzing consumer demand for *large* numbers of products. We allow the consumer to purchase multiple units of any product and to purchase multiple products at once (think of a consumer selecting a bundle of goods in a supermarket). In our model each product has an associated unobservable vector of attributes from which the consumer derives utility. Our model allows for heterogeneous utility functions across consumers, complex patterns of substitution and complementarity across products, and nonlinear price effects. The dimension of the attribute space is, by assumption, much smaller than the number of products, which effectively reduces the size of the consumption space and simplifies estimation. Nonetheless, because the number of bundles available is massive, a new estimation technique, which is based on the practice of negative sampling in machine learning, is needed to sidestep an intractable likelihood function. We prove consistency of our estimator, validate the consistency result through simulation exercises, and estimate our model using supermarket scanner data.

JEL classification: C13, C34, D12, L20, L66

Keywords: discrete choice, demand estimation, negative sampling, machine learning, scanner data

Acknowledgements: We are grateful to participants at the NBER's 2022 Summer Institute Industrial Organization Workshop for several encouraging comments and suggestions. We also acknowledge Emmet Hall-Hoffarth for excellent research and software development assistance.

*China Center for Behavioral Economics and Finance, Southwestern University of Finance and Economics, Chengdu

†St Hugh's College and the Department of Economics, University of Oxford

‡Department of Economics, Johns Hopkins University

1 Introduction

Knowing what consumers will purchase in different circumstances is crucial for answering many economic and business questions. What is often needed is not merely the own-price elasticity of a single good but rather knowledge of the full demand system of a large swathe of interrelated products. For instance, a government instituting a tax on sodas must know the health content of each product into which consumers substitute to calculate the health effects of the policy. A government instituting a new trade policy must know the change in demand for each complement good and substitute good for the products directly affected by the policy in order to calculate changes in welfare. A firm entering a new market must understand how each of its products interacts with the existing products in the market in order to know which goods to introduce and at which prices.

This crucial knowledge of consumer behavior can be difficult to acquire because of the enormous number of interconnected goods in any modern economy. This difficulty can be overcome by (i) using institutional knowledge to handpick a small subset of goods to analyze and (ii) aggregating many products into a few broad categories of goods.[1] Relying on these procedures, economists often analyze demand systems containing only a small number of goods.[2,3] While economic theory provides conditions under which this focused analysis is justified it is generally recognized that these conditions are very stringent.[4] In principle the demand for any good is a function of the prices of all other goods. Thus, when the analysis is confined to only a few goods it seems likely that omitted variable bias is introduced from excluding relevant prices or aggregating the prices of goods into

[1]Many studies reduce the number of goods by focusing on a few "inside" goods and treating all other goods as a single "outside" good. This approach is best understood as an instance of procedure (ii). In particular, the restrictive assumptions required to justify (ii) need to hold for the outside good.

[2]For instance, Lewbel and Pendakur (2009) estimate a demand system with 9 goods, Hausman and Newey (2016) provide bounds on the equivalent variation of a price change with unrestricted preference heterogeneity with an application containing 2 goods (gasoline and an "other" good), Blundell et al. (2017) show how to impose a consequence of economic theory onto a quantile regression estimator and provide an application with 2 goods (again, gasoline and an "other" good), and Kitamura and Stoye (2018) test household expenditure data for consistency with a non-parametric random utility model using between 3 and 5 goods. An exception is discrete choice models where goods have observable attributes. See the "Related approaches" section below.

[3]One reason for this fact is that many models and estimation techniques used by economists become untenable when the number of goods is too large. See Berry et al. (2014) and Keane (2015) for a discussion of some of the issues.

[4]See Deaton and Muellbauer (1980), Blundell and Robin (2000), and, for a more recent discussion, see Chernozhukov et al. (2019).

a single price index. As our approach is suitable for estimating large demand systems we believe it has the potential to greatly reduce this omitted variable bias.

We present a tractable random utility model and estimation method, inspired by recent innovations in machine learning, which is suitable for analyzing panel datasets with thousands of products and hundreds of thousands of purchases (for instance, supermarket scanner data). In our model each product has an associated unobservable (to the analyst) vector of attributes. The consumer purchases a bundle of products to maximize utility which is derived from the attributes of the products. The attribute space is assumed to be much smaller than the number of products. This greatly diminishes the effective size of the consumption space and greatly reduces the number of parameters which need to be estimated. Despite this parsimony, we demonstrate that our model has flexible price effects, and in particular, goods can be substitutes, complements, or independent.

The parameters of our model cannot be estimated by straightforward maximum likelihood methods because the likelihood function is prohibitively difficult to calculate for applications where the number of alternatives is large (note that in our model an alternative is a bundle of goods). This intractability motivates our introduction of a new estimation technique, inspired by the idea of negative sampling from machine learning, which is suitable for estimating parameters in a discrete choice model with a large choice space. The basic idea of the technique is to perform maximum likelihood estimation conditioning on simulated random variables. Through carefully designing the distribution of the simulated random variables we are able to obtain a tractable conditional likelihood function even though the original likelihood function is intractable.

1.1 Our model

Suppose there are L goods and each good $\ell \leq L$ has some unobserved vector of attributes $\boldsymbol{\alpha}_\ell \in \mathbb{R}^K$. We assemble the L attribute vectors into an attribute matrix $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K]$. The consumer purchases a bundle of goods given by a vector $\boldsymbol{q} = [q_1, \ldots, q_L]$ where each q_ℓ is a non-negative integer. For instance $\boldsymbol{q} = [1, 0, 3]$ denotes a bundle containing one unit of good 1, zero units of good 2, and three of good 3. Each bundle $\boldsymbol{q}$ has an associated vector of attributes given by $\boldsymbol{Aq} = \sum_{\ell=1}^{L} q_\ell \boldsymbol{\alpha}_\ell$. Each good has a price p_ℓ which is assembled into a price vector $\boldsymbol{p} = [p_1, \ldots, p_L]$. The utility function of a consumer is a function of the attributes consumed, $\boldsymbol{Aq}$, the amount of money spent,

$\boldsymbol{p}'\boldsymbol{q}$, and a random shock, and is given by

$$U(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{q} - d\boldsymbol{p}'\boldsymbol{q} + \varepsilon_{\boldsymbol{q}} \tag{1}$$

where $\boldsymbol{b} \in \mathbb{R}^K$, $\boldsymbol{A}$, and $d > 0$ are parameters and $\varepsilon_{\boldsymbol{q}}$ is an i.i.d. standard Gumbel shock. While this utility function is quasilinear the utility function we present in the main paper can have more flexible income effects. In our application $\boldsymbol{b}$ and d are consumer specific while the attributes in $\boldsymbol{A}$ are constant across consumers.

The utility function in equation (1), because of the Gumbel shocks, is an instance of a multinomial logit model where an alternative is a *bundle* of goods in contrast to the usual setup where an alternative would be a single unit of a single good. Given the well-known results on the multinomial logit model (see McFadden (1974)) it is not surprising that the probability with which the consumer with the utility function in equation (1) purchases bundle $\boldsymbol{q}$ when prices are $\boldsymbol{p}$ is given by

$$f(\boldsymbol{q}|\boldsymbol{p}) = \frac{\exp(U(\boldsymbol{q}, \boldsymbol{p}))}{\sum_{\tilde{\boldsymbol{q}}} \exp(U(\tilde{\boldsymbol{q}}, \boldsymbol{p}))} \tag{2}$$

where the summands iterates over all consumption bundles. The multinomial logit model is known to display unrealistic price effects in the usual case when each alternative is a single unit of a single good. However, this criticism no longer applies when each alternative is a consumption bundle (even when utility is quasilinear). The reason is that a price change of a particular good impacts the desirability of all bundles in which the good appears. Thus a price change for a single product alters the desirability of many bundles. This means that the impact of the price change for other products can display complicated patterns; in particular, it could depend on the extent to which they are co-purchased with the product whose price has changed.

As a simple example of our model, consider a customer in a teashop, selecting a bundle to consume. Three goods are on offer: tea, biscuits, and cake. Unlike our general model (which allows for the consumption of multiple units of each good), let us suppose that a maximum of just one unit of each good can be consumed. This gives rise to $2^3 = 8$ bundles which are listed in Table 1. Formally, each bundle can be represented by a vector $\boldsymbol{q} \in \{0, 1\}^3$. Let us say that our tearoom is offering 'warm', 'sweet' and 'filling' consumption with its products, with each good delivering these attributes in different

Bundles	**tea**	**biscuits**	**cake**	**net utility**	**demand probability**	**Comments**
nothing	0	0	0	0.0	0%	no utility
pot of tea	1	0	0	0.3	0%	just a drink
biscuits	0	1	0	2.8	1%	slightly desirable
slice of cake	0	0	1	5.7	26%	pretty good
tea and biscuits	1	1	0	4.0	5%	somewhat desirable
tea and cake	1	0	1	6.1	36%	just the ticket
cake and biscuits	0	1	1	4.3	7%	thirsty work
high tea	1	1	1	5.7	25%	a bit much

Table 1: List of bundles and their demand probabilities.

quantities. This is captured by the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0.5 \\ -0.5 & 1 & 1.1 \\ 0 & 1.2 & 0.8 \end{bmatrix}.$$

where the first column is the attribute vector for tea, the second for biscuits, and the third for cake. We allow a good to have a negative amount of some attribute; in particular, tea has 1 unit of the 'warm' attribute, nothing of 'filling', and -0.5 units of 'sweet'.

The net utilities reported in Table 1 are calculated using equation (1), with $\boldsymbol{b} = (5, 5, 1)$, $d = 1$, and the price of each good set at 1 (so $\boldsymbol{p} = (1, 1, 1)$). Net utilities in turn give rise to bundle-demand probabilities (via (2)). From these, the demand for goods can be calculated. For example, as biscuits appear in exactly the bundles 'Biscuits' (prob. .01), 'Tea and Biscuits' (.05), 'Cake and Biscuits' (.07), and 'High Tea' (.25), its expected demand is 0.01+0.05+0.07+0.25 = 0.38.

1.2 Our estimation technique

Maximum likelihood methods are not suitable for estimating the parameters in our model when the number of goods is large. The reason is that the denominator in the probability mass function given by (2) has an infinite sum which is impossible to approximate quickly when there are too many goods. We introduce an alternative way to estimate parameters. To make the discussion more concrete let $\boldsymbol{c}(\boldsymbol{p})$ be a random consumption bundle with probability mass function $f(\boldsymbol{q}|\boldsymbol{p})$ given by (2).

The problem with using maximum likelihood estimation is that the support of $\boldsymbol{c}(\boldsymbol{p})$ is too large, that is, there are too many possible bundles. Our solution is to define a new

random variable $\mathcal{S}$ so that the distribution of $\boldsymbol{c}(\boldsymbol{p})$ conditional on $\mathcal{S}$ has a much smaller support. More specifically, let $S(\boldsymbol{q})$ be a random set containing a small list of bundles, one of which is the bundle $\boldsymbol{q}$. Intuitively, $S(\boldsymbol{q})$ is giving a hint about what bundle it was passed. Now, let $\mathcal{S} = S(\boldsymbol{c}(\boldsymbol{p}))$ and let $f(\boldsymbol{q}|Q, \boldsymbol{p})$ be the probability which the consumer with the utility function given by (1) purchases $\boldsymbol{q}$ conditional on $\mathcal{S} = Q$. Provided $S(\boldsymbol{q})$ satisfies certain conditions we show

$$f(\boldsymbol{q}|Q, \boldsymbol{p}) = \frac{\exp(U(\boldsymbol{q}, \boldsymbol{p}))}{\sum_{\tilde{\boldsymbol{q}} \in Q} \exp(U(\tilde{\boldsymbol{q}}, \boldsymbol{p}))} \tag{3}$$

Note that the denominator in (3) only has as many terms as are contained in the set Q. Thus, provided $S(\boldsymbol{q})$ is constructed in a way so that it outputs smaller sets Q we are guaranteed a tractable conditional probability mass function and so conditional maximum likelihood estimation is also tractable. We show in the body of the paper that this conditional maximum likelihood estimator is consistent.

The idea behind this estimation technique comes from negative sampling introduced by Mikolov et al. (2013). They recommend using a probability mass function similar to (3) to estimate parameters but do not provide much in the way of theoretical justification for the approach. In contrast, we provide conditions under which the probability mass function of $\boldsymbol{c}(\boldsymbol{p})$ conditional on $\mathcal{S}$ is given by equation (3) and show that maximizing the corresponding likelihood function yields a consistent estimator of the parameters.

1.3 Related approaches

There is a large literature on estimating demand in economics. See the reviews of Barnett and Serletis (2008) on demand for continuous commodities, Greene (2015) for discrete choice models using panel data, and Berry and Haile (2021) and Gandhi and Nevo (2021) for discrete choice models used in industrial organization.

While the continuous commodities literature almost exclusively analyzes small numbers of goods the discrete choice literature has tackled larger demand systems by assuming that consumers derive utility from the attributes of products. Suppose there are L products and each product analyzed has an associated K vector of observable attributes. Assuming that utility is derived from attributes and provided K is small the consumption space is effectively K dimensional instead of L. Examples where utility is derived from observable attributes are Berry et al. (1995), Hendel (1999), Nevo (2001),

Berry et al. (2004), and Dubois et al. (2020). What distinguishes our model from this literature is that in our model the attributes of products are unobserved.[5] This means that we have more parameters to estimate which, on the one hand makes the estimation more challenging but on the other hand introduces greater flexibility into our model.

Machine learning techniques are designed for large scale applications. Thus, it seems natural to look at machine learning for a way to estimate large demand systems. The potential that machine learning holds for economics is discussed in Varian (2014), Mullainathan and Spiess (2017), and Athey and Imbens (2019). Indeed, machine learning techniques have been used to estimate demand in Bajari et al. (2015), Wan et al. (2017), and Ruiz et al. (2020). Bajari et al. (2015) compare out-of-sample fit of predicted demand between different machine learning methods (they also fit and compare a simple linear and logistic regression as well). Wan et al. (2017) estimate a simple three step model of consumer behavior where each consumer first decides which categories of goods to buy, then decides which product to buy in a category, and finally decides on the quantity of the product. Ruiz et al. (2020) build and estimate a model inspired by the machine learning sub-field of natural language processing. Their model assumes that a consumer constructs a consumption bundle sequentially. That is, products are added to the consumption bundle one-by-one to maximize the random utility of the currently assembled bundle and not the utility of the actual bundle to be purchased (they also present an extension where the consumer can think one item ahead). What distinguishes our approach from these studies is our tight connection to standard economic theory. We use a random utility model built on the assumption that utility is derived from the bundle actually purchased. Because our consumers are utility maximizers we can credibly use our model to discuss welfare issues.

1.4 Remainder of the paper

In the next Section we set out the model in full, discuss parameter identification, present some observations and comparative statics results regarding price complementarity, and outline our approach to the price endogeneity problem. In Section 3 we provide formal results on consistent estimation. Section 4 develops a computationally effective estimation algorithm based on these results, and simulates its properties. The algorithm is

[5] It is easy to extend our model to include both observed and unobserved attributes.

applied using real data from the market research company, DunnHumby, in Section 5.

2 A Random Utility Model

In this section we introduce our random utility model which is suitable for analyzing demand for consumption bundles.

2.1 The Model Defined

Let $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$ denote the non-negative integers. There are L goods (call them good 1, good 2, ..., good L) which can only be consumed in non-negative integer quantities. A vector $\boldsymbol{q} = (q_1, q_2, \ldots, q_L) \in \mathbb{N}_0^L$ is a consumption bundle where $q_\ell \in \mathbb{N}_0$ is the amount of good ℓ consumed. Each good, $\ell \in \{1, \ldots, L\}$, has an associated unobservable K-vector of attributes $\boldsymbol{\alpha}_\ell \in \mathbb{R}^K$ where $K \leq L$. The L attribute vectors are placed into a $K \times L$ matrix $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_L]$ called an *attribute matrix*.

Each good ℓ has a price p_ℓ which we place into a price vector $\boldsymbol{p} = [p_1, \ldots, p_L] \in \mathbb{R}_{++}^L$. The deterministic utility (a random shock will be added later) of bundle $\boldsymbol{q}$ when prices are $\boldsymbol{p}$ takes the following quadratic form

$$U(\boldsymbol{q}, \boldsymbol{p}) = \underbrace{\boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{q}}_{\text{utility from attributes}} - \underbrace{d_1\boldsymbol{p}'\boldsymbol{q} - d_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2d_3(\boldsymbol{A}_1'\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})}_{\text{dis-utility from expenditure}} \tag{4}$$

where $\boldsymbol{A}$ is the $K \times L$ attribute matrix, $\boldsymbol{A}_1 \in \mathbb{R}_+^K$ is the first row of $\boldsymbol{A}$, $\boldsymbol{b} \in \mathbb{R}^K$, and $d_1 > 0, d_2 \geq 0$, and $d_3 \geq 0$. A utility function satisfying (4) is called a *standard quadratic in unobserved attributes* (standard Qua) utility function.[6]

A consumer with a standard Qua utility function wants to purchase consumption bundles with good attributes at a reasonable price. In other words, they like bundles with attractive attributes (captured by the "utility from attributes" term) and dislike spending money (captured by the "dis-utility from expenditure" term).

Inspection of the "dis-utility from expenditure" term in equation (4) reveals that, in several ways, the standard Qua utility function gives the first unobserved attribute, $\boldsymbol{A}_1$, a special role. First, only the first attribute even enters this "dis-utility from expenditure" term. Second, by assuming that $\boldsymbol{A}_1 \in \mathbb{R}_+^K$ we are requiring every good to have a non-

[6]Throughout we maintain the convention that $\boldsymbol{\alpha}_\ell$ is column ℓ of $\boldsymbol{A}$ while $\boldsymbol{A}_k$ is row k of $\boldsymbol{A}$.

negative quantity of attribute 1.[7] Indeed, we allow goods to possess negative quantities of any attribute other than attribute 1. For instance, the attribute vector for bananas might be $\boldsymbol{\alpha}_{\text{bananas}} = [3, -12, -5.3]$. Although the special role given to attribute 1 may seem restrictive this is done without loss of generality in a sense discussed in Section 2.2 and made formal in Proposition 2 below.

The utility function in equation (4) is quasi-linear when $d_2 = d_3 = 0$. While quasi-linear utility may be sensible in some instances there may be cases where it is important to penalize large purchases (which can be captured by the quadratic term $d_2(\boldsymbol{p}'\boldsymbol{q})^2$) and there may be cases where the dis-utility of expenditure is product specific (which can be captured by the term $2d_3(\boldsymbol{A}_1'\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})$).[8]

We use our model to analyze supermarket scanner data which has repeated purchases for each consumer. We assume $\boldsymbol{A}$ is constant across consumers but allow $\boldsymbol{b}, d_1, d_2$, and d_3 to vary by consumer. In much of the theoretical analysis we treat $\boldsymbol{b}, d_1, d_2$, and d_3 as constants and so these theoretical results pertain to the behavior of a single consumer with constant parameter values.

In our model consumers are random utility maximizers and so purchase consumption bundles to maximize

$$\tilde{U}(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}, \boldsymbol{p}) + \varepsilon_{\boldsymbol{q}}. \tag{5}$$

where U is a standard Qua utility function and $\varepsilon_{\boldsymbol{q}}$ is an iid standard Gumbel shock. The random utility function $\tilde{U}$ is called a *standard Qua random utility function*. Our model of the individual consumer is therefore a standard multinomial logit model where each alternative is a consumption bundle. The following proposition reports the probabilities with which our consumer purchases different bundles.

Proposition 1. *Let $\tilde{U}$ be a standard Qua random utility function, let $\boldsymbol{p} \in \mathbb{R}^L_{++}$, and let*

[7] The requirement that $\boldsymbol{A}_1 \geq \boldsymbol{0}$ is essential for ensuring that the utility function is strictly decreasing in expenditure. In fact, the restrictions imposed on the parameters ensure that U is strictly concave in attributes (i.e. in $\boldsymbol{A}\boldsymbol{q}$) and concave and decreasing in expenditure. These properties are needed for Proposition 1. In particular, without these restrictions we cannot ensure that the denominator in (7) is not ∞.

[8] When utility is quasi-linear (and so $d_2 = d_3 = 0$) then, using standard arguments, the utility function in (4) can be derived from a more primitive model wherein price does not directly enter the utility function but is incorporated through a budget constraint. This derivation does not work when $d_2 > 0$ or $d_3 > 0$ and so, in general, the standard Qua utility function takes as a primitive assumption that the consumer loses utility from spending money. This type of model, called an *expenditure augmented* utility function, is introduced and studied in Deb et al. (Forthcoming).

$f(\cdot|\boldsymbol{p}) : \mathbb{N}_0^L \to [0,1]$ *be defined by*

$$f(\boldsymbol{q}|\boldsymbol{p}) = P\left(\tilde{U}(\boldsymbol{q},\boldsymbol{p}) \geq \sup_{\tilde{\boldsymbol{q}}\in\mathbb{N}_0^L} \tilde{U}(\tilde{\boldsymbol{q}},\boldsymbol{p})\right), \qquad \textit{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{6}$$

Then,

$$f(\boldsymbol{q}|\boldsymbol{p}) = \frac{\exp(U(\boldsymbol{q},\boldsymbol{p}))}{\sum_{\tilde{\boldsymbol{q}}\in\mathbb{N}_0^L}\exp(U(\tilde{\boldsymbol{q}},\boldsymbol{p}))} \tag{7}$$

The probability mass function $f(\boldsymbol{q}|\boldsymbol{p})$ defined in (6) is the probability with which the consumer purchases $\boldsymbol{q}$ when prices are $\boldsymbol{p}$. Equation (7) is not surprising given the classic results on the multinomial logit model (see McFadden (1974)). In fact, the only difficulty in the proof is showing that the denominator in (7) is finite.

2.2 A Seemingly More General Model

There are two features of the standard Qua utility function which appear restrictive but turn out not to be. First, letting $[a_1,\ldots,a_K] = \boldsymbol{A}\boldsymbol{q}$ denote the attribute vector obtained by purchasing $\boldsymbol{q}$, we see that the quadratic expression in (4), namely $\boldsymbol{q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{q} = \sum_{k=1}^K a_k^2$, does not allow different attributes to interact multiplicatively. One might think that the more general quadratic expression $\boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{q} = \sum_{k=1}^K\sum_{k'=1}^K B_{k,k'}a_k a_{k'}$ is preferable (here $\boldsymbol{B}$ is a $K\times K$ positive definite matrix with entries $B_{k,k'}$). Second, the expression $2d_3(\boldsymbol{A}_1'\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})$ appears to give attribute 1 a special and unjustified role. One might think it preferable to replace this term with $-2(\tilde{\boldsymbol{d}}'\boldsymbol{A}\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})$ as it does not privilege any of the attributes. It turns out that both of these seeming improvements to the standard Qua model result in exactly the same class of utility functions.

Specifically, the following class of utility function, although incorporating many more parameters, contains exactly the same utility functions (and thus can describe the same behavior) as the standard Qua class. This class, whose members are called *general quadratic in unobserved attributes (general Qua) utility functions*, is defined by

$$U(\boldsymbol{q},\boldsymbol{p}) = \underbrace{\boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{q}}_{\text{utility from attributes}} - \underbrace{d_1\boldsymbol{p}'\boldsymbol{q} - d_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2(\tilde{\boldsymbol{d}}'\boldsymbol{A}\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})}_{\text{dis-utility from expenditure}}$$

where $\boldsymbol{A}$ is a $K\times L$ attributes matrix, $\boldsymbol{b}\in\mathbb{R}^K$, $\boldsymbol{B}$ is a $K\times K$ positive definite matrix, $d_1 > 0$, $d_2 \geq 0$, and $\tilde{\boldsymbol{d}}\in\mathbb{R}^K$ where $\tilde{\boldsymbol{d}}$ is restricted so that $\tilde{\boldsymbol{d}}'\boldsymbol{A}\boldsymbol{q} \geq 0$ for all $\boldsymbol{q}\in\mathbb{N}_0$.[9] The

[9]The restrictions placed on the parameters (such as $\tilde{\boldsymbol{d}}'\boldsymbol{A}\boldsymbol{q} \geq 0$) ensure that U is strictly concave in attributes (i.e. $\boldsymbol{A}\boldsymbol{q}$) and concave and strictly decreasing in expenditure. As mentioned earlier, these properties are needed to ensure that Proposition 1 holds. Without them, we cannot ensure that the denominator in (7) is finite.

general Qua class of utility functions appears to improve on the standard Qua class in the ways suggested earlier and yet the following proposition shows that the two classes are the same.

Proposition 2. *For every general Qua utility function U with a $K \times L$ attribute matrix $\boldsymbol{A}$ there exists a standard Qua utility function $\tilde{U}$ which also has a $K \times L$ attribute matrix $\tilde{\boldsymbol{A}}$ so that*

$$U(\boldsymbol{q}, \boldsymbol{p}) = \tilde{U}(\boldsymbol{q}, \boldsymbol{p}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \text{ and all } \boldsymbol{p} \in \mathbb{R}_{++}^L$$

The message of Proposition 2 is that every general Qua utility function is also a standard Qua utility function. Of course, the converse is also true. In the rest of the paper we focus our analysis on consumers with standard Qua utility functions but, given Proposition 2, we could carry out all our analysis on consumer's with general Qua utility functions and none of the results would change.

2.3 Identification

We discuss identification of the parameters of our model given consumption data. Let $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3]$ denote a list of the parameters of the standard Qua model. Let Θ denote all such lists where we assume that there are always K unobserved attributes (and so, $\boldsymbol{A}$ is always a $K \times L$ matrix.). Let $U(\cdot; \boldsymbol{\theta})$ denote the standard Qua utility function with parameter values $\boldsymbol{\theta}$. Let $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta})$ denote the stochastic choice function generated by $U(\cdot; \boldsymbol{\theta})$ in the sense of (6).

Proposition 3. *Define a function ψ with domain Θ by*

$$\psi([\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3]) = [\boldsymbol{A}'\boldsymbol{A}, \boldsymbol{A}'\boldsymbol{b}, d_1, d_2, d_3\boldsymbol{A}_1] \tag{8}$$

where $\boldsymbol{A}_1$ is the first row of $\boldsymbol{A}$. Let $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$ and let $E \subseteq \mathbb{R}_{++}^L$ be a non-empty open set. The following are equivalent.

1. $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}) = f(\boldsymbol{q}|\boldsymbol{p}; \tilde{\boldsymbol{\theta}}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \text{ and all } \boldsymbol{p} \in E.$
2. $U(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\theta}) = U(\boldsymbol{q}, \boldsymbol{p}; \tilde{\boldsymbol{\theta}}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \text{ and all } \boldsymbol{p} \in E.$
3. $\psi(\boldsymbol{\theta}) = \psi(\tilde{\boldsymbol{\theta}}).$

Proposition 3 says that two parameter vectors $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$ are empirically indistinguishable if and only if $\psi(\boldsymbol{\theta}) = \psi(\tilde{\boldsymbol{\theta}})$.

2.4 Interpreting Attribute Parameters

From Proposition 3 it is clear that the bulk of the content of the attribute matrix $\boldsymbol{A}$ which can be learned from consumption data is $\boldsymbol{A}'\boldsymbol{A}$. In other words, all we can learn are the dot product terms $\boldsymbol{\alpha}_\ell'\boldsymbol{\alpha}_j$ for each $\ell, j \in \{1, \ldots, L\}$. How should we interpret these dot products? To answer this, suppose that $\boldsymbol{c}(\boldsymbol{p})$ is a random consumption bundle in $\mathbb{N}_0^L$ with distribution $f(\boldsymbol{q}|\boldsymbol{p})$ given by (7). Let ℓ and j be two distinct goods of interest. Let $E_{\ell,j}$ be the event that the consumer purchases a positive quantity of both goods ℓ and j, let E_ℓ be the event that the consumer purchases a positive quantity of good ℓ and no units of good j and let E_0 be the event that the consumer purchases no units of good ℓ and no units of good j. It is easy to see, using equation (7), that $P(E_{\ell,j})/P(E_\ell)$ is strictly decreasing in $\boldsymbol{\alpha}_\ell'\boldsymbol{\alpha}_j$ while $P(E_\ell)/P(E_0)$ is unaffected by the value of $\boldsymbol{\alpha}_\ell'\boldsymbol{\alpha}_j$. In other words, ceteris paribus lower values of $\boldsymbol{\alpha}_\ell'\boldsymbol{\alpha}_j$ mean that a consumer is more likely to co-purchase goods ℓ and j rather than purchasing only ℓ or only j. On the other hand, higher values of $\boldsymbol{\alpha}_\ell'\boldsymbol{\alpha}_j$ do not make it any more or less likely that a consumer purchases some of good ℓ (and no good j) compared to purchasing no units of good ℓ.

2.5 Price Effects

Here we consider the price effects for an individual consumer with a standard Qua random utility function. Our model of the individual consumer is essentially an instance of a multinomial logit model where alternatives are bundles. The multinomial logit model is commonly believed to overly restrict price effects. While this point seems incontrovertible when a consumer can only purchase a single unit of a single good we shall show that the price effects in our model are more flexible.

We begin by reiterating the point that price effects in the logit model are very restricted when the consumer may only purchase a single unit of a single product. So, consider a consumer with a standard Qua random utility function with quasilinear utility (so $d_2 = d_3 = 0$) and suppose that the consumer can only purchase, at most, one unit of a single product. Fix some price vector $\boldsymbol{p} \in \mathbb{R}_{++}^L$ and, for each ℓ, let π_ℓ be the probability with which the consumer purchases good ℓ. Under these assumptions it is easy to show

$$\frac{\partial \pi_j}{\partial p_\ell} = d_1 \pi_j \pi_\ell, \qquad \text{for all } j \neq \ell \tag{9}$$

The price effects described by equation (9) are indeed very restricted. In particular,

all goods are substitutes and the intensity of substitution is determined by the single parameter d_1 as well as the probability of consuming each good.

The following proposition derives price-effects when the consumer is not restricted to purchasing a single unit of a single good.

Proposition 4. *Let* $\boldsymbol{c}(\boldsymbol{p}) = [c_1(\boldsymbol{p}), c_2(\boldsymbol{p}), \ldots, c_L(\boldsymbol{p})]$ *be a random consumption bundle whose probability mass function* $f(\boldsymbol{q}|\boldsymbol{p})$ *satisfies* (7) *where* U *is a standard Qua utility function with* $d_2 = d_3 = 0$. *Then,*

$$\partial_{\boldsymbol{p}} \boldsymbol{E}[\boldsymbol{c}(\boldsymbol{p})] = -d_1 \, \mathbf{Var}\left(\boldsymbol{c}(\boldsymbol{p})\right) \tag{10}$$

Let us rewrite equation (10) in a way which is more comparable with (9). To this end, for each ℓ, let $\pi_\ell = \boldsymbol{E}[c_\ell(\boldsymbol{p})]$ and let $\pi_{j,\ell} = \boldsymbol{E}[c_j(\boldsymbol{p})c_\ell(\boldsymbol{p})]$. Now, equation (10) implies

$$\frac{\partial \pi_j}{\partial p_\ell} = d_1 \pi_j \pi_\ell - d_1 \pi_{j,\ell}, \qquad \text{for all } j \neq \ell \tag{11}$$

There is a very intuitive story behind the price effects described by (11). A rise in the price of good ℓ makes every bundle which contains some units of good ℓ more expensive and so decreases their probability of being purchased. On the other hand, every bundle in which no units of ℓ are consumed is relatively more attractive. As a result, whether the demand for good j rises or falls depends on whether it is commonly purchased with good ℓ or not. That is, the crucial question for determining the direction of price effects is whether j often appears alongside ℓ or not. This tendency to be co-purchased is captured by the $\pi_{j,\ell}$ expression.[10]

We mention two ways in which the price effects described by our model in equation (11) are more flexible than those of the single-product single-unit model in (9). First, in (11) goods may complements (just consider a large $\pi_{j,\ell}$ term); a property ruled out by equation (9). Second, in (11) we may have $\pi_1 = \pi_2$ without requiring $\partial \pi_1 / \partial p_3 = \partial \pi_2 / \partial p_3$. In other words, the demand for two different goods which have similar probabilities of being purchased needn't have the same response to the change in the price of some third good. Again, this is ruled out by (9).

[10] Gentzkow (2007) warns that inferring two good are complements based on the fact that they are often co-purchased in a dataset with multiple consumers can be highly problematic. The problem is that the observation that two goods are often either purchased together or not at all can both be explained either as complementarity or as a particular type of preference heterogeneity. This problem clearly affects the analysis of cross sectional data but, as we are considering the demand of a single consumer, there is no danger of confusing preference heterogeneity with complementarity. Recall that our application uses panel data where we allow the parameters $\boldsymbol{b}, d_1, d_2$ and d_3 to vary by consumer.

Note that Proposition 4 requires $d_2 = d_3 = 0$ which is an assumption we do not impose in our empirical application. The following proposition describes price effects in the more general case.

Proposition 5. *Let* $\boldsymbol{c}(\boldsymbol{p}) = [c_1(\boldsymbol{p}), c_2(\boldsymbol{p}), \ldots, c_L(\boldsymbol{p})]$ *be a random vector whose probability mass function* $f(\boldsymbol{q}|\boldsymbol{p})$ *satisfies* (7) *where* U *is a standard Qua function. Then,*

$$\partial_{\boldsymbol{p}} \boldsymbol{E}[\boldsymbol{c}(\boldsymbol{p})] = \mathbf{Cov}\left(\boldsymbol{c}(\boldsymbol{p}), \partial_{\boldsymbol{p}} U(\boldsymbol{c}(\boldsymbol{p}), \boldsymbol{p})\right) \tag{12}$$

Proposition 4 is a straightforward corollary of Proposition 5. We have argued that the price effects described by (10) are not overly restricted. The price effects described by (12) are of course even more flexible and when incorporating preference heterogeneity, as we do in the application, the price effects will be greater still. One can get a deeper understanding of the price effects implied by (12) by plugging in the definition of $U(\boldsymbol{c}, \boldsymbol{p})$ which yields an expression with the parameters d_1, d_2, d_3, and $\boldsymbol{A}_1$.

2.6 Price Endogeneity and Period Dummies

The demand function of a consumer can change over time due to changes in preferences (for instance, in response to an advertising campaign) or due to changes in the quality of goods (for instance, in-season strawberries are tastier than out-of-season strawberries). It is important to account for these shifts in the demand function because these shifts can make it hard to distinguish whether demand for a good has changed due to price movements or if prices are moving in response to a change in the demand function. This lack of clarity in the direction of causation between price and demand can confound the estimation of preference parameters. To tackle this problem we explicitly model preference / quality change.

Similar to the method employed in Ruiz et al. (2020) our approach is to add period specific dummy variables to the utility function. These dummy variables shift utility each period (in our application a period is 4 weeks) and so they can capture shifts in utility arising from taste change or quality change. More concretely, suppose we have some purchasing data from a consumer for time periods between T_0 and T_1. We partition the interval $(T_0, T_1] \subseteq \mathbb{R}$ into T^* sub-intervals of equal length, $\tau_1, \ldots, \tau_{T^*}$ (so each τ_k is a interval in $\mathbb{R}$) and assume that when the consumer goes shopping at time period $t \in \tau_k$

they evaluate bundles according to the utility function $\tilde{U}_{\tau_k}$ defined by

$$\tilde{U}_{\tau_j}(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{b}'_{\tau_j} \boldsymbol{A}\boldsymbol{q} + \tilde{U}(\boldsymbol{q}, \boldsymbol{p}) \tag{13}$$

where $\tilde{U}$ is the standard Qua random utility function of equation (5) while $\boldsymbol{b}_{\tau_j} \in \mathbb{R}^K$ is a period specific term which shifts the utility of the attributes.

In equation (13) entry number k of $\boldsymbol{b}_{\tau_j}$ represents a shift to the utility received from attribute k. Note that we are assuming that all changes to preferences and quality can be represented by changes to the desirability of the latent attributes. For instance, perhaps attribute k correlates highly with fruits which are in-season in the summer. If this is the case then our model can pick up changes in the quality of these fruits by letting entry k of $\boldsymbol{b}_{\tau_j}$ take large values for τ_j in the summer and otherwise $\boldsymbol{b}_{\tau_j}$ takes smaller values.

Recall that we allow the parameters $\boldsymbol{b}$ and d_1, d_2, and d_3 to vary by consumer in our empirical section. By contrast, we envisage that the period-specific dummies $\boldsymbol{b}_{\tau_j}$ are held constant across consumers.

3 Estimation

Although our application uses panel data from many consumer we shall first consider how to estimate the parameters of our model given time series data from a single consumer. For now, we also assume that there are no period-specific shifts to the utility function (as in Equation (13)). We proceed in this way to keep the exposition and notation simple. Our method generalizes to the panel data with period-specific dummy context in a straightforward manner. When dealing with panel data we allow that each consumer has individual specific parameters $\boldsymbol{b}, d_1, d_2$ and d_3. Thus, only the attribute matrix $\boldsymbol{A}$ is assumed common to all consumers.[11]

We cannot and do not attempt to estimate parameters by maximizing the conditional log likelihood of the data. The reason is that our underlying probability mass function

[11]This could lead to an incidental parameters problem if the number of consumers in the dataset is much larger than the number of observations per consumer. That is, as each new consumer in the dataset adds an additional $K + 3$ parameters to be estimated, a data set with too many consumers relative to observations per consumer can involve too many parameters to be estimated reliably. Recently, Dubois et al. (2020) have estimated demand for soda using panel data assuming, in a similar fashion to our approach, consumer-specific parameters. They provide evidence that the incidental parameters problem does not hinder the quality of their estimates due to the fact that they have a large number of observations per consumer.

(7) is intractable. Specifically, the denominator in (7) has too many terms.[12] Instead, we define a random function S, called a signal function, which assists our estimation. Specifically, we construct the signal function in a way so that probability mass function of demand conditional on S takes a particularly tractable form.

To make our discussion more concrete suppose that $\boldsymbol{c}(\boldsymbol{p})$ is a random consumption bundle with a probability mass function $f(\boldsymbol{q}|\boldsymbol{p})$ given by (7). Let $S(\boldsymbol{q})$ be a random set which contains $\boldsymbol{q}$ (along with other bundles). Intuitively, $S(\boldsymbol{q})$ is providing a (smallish) list of bundles, one of which is the correct bundle while the rest are imposters. Now, provided S satisfies certain conditions, the probability mass function of $\boldsymbol{c}(\boldsymbol{p})$ conditional on $S(\boldsymbol{c}(\boldsymbol{p}))$ will be the same as the probability mass function of the consumer who was constrained to select from a bundle in $S(\boldsymbol{c}(\boldsymbol{p}))$. In other words, implementing the signal function method allows us to treat the consumption space as if it is $S(\boldsymbol{c}(\boldsymbol{p}))$ instead of the much larger $\mathbb{N}_0^L$.

When designing S there is a trade-off between tractability and efficiency. Intuitively, as more information is passed through $S(\boldsymbol{c}(\boldsymbol{p}))$ the estimator becomes less efficient yet easier to calculate. It is well known that conditional MLE is generally less efficient than unconditional MLE and that the greater the set of conditional variables the less efficient is the estimator (see Property 7.18 of maximum likelihood estimators in Section 7.5.3 of Gourieroux and Monfort (1995)). The same idea is at work with the signal function. An $S(\boldsymbol{c}(\boldsymbol{p}))$ which gives too much information is akin to performing MLE while conditioning on too many variables.

It may be instructive to consider two extreme cases of the signal function. First, suppose for all $\boldsymbol{q}$ we have $S(\boldsymbol{q}) = \mathbb{N}_0^L$. Now, the signal function passes no information. Unconditional MLE and MLE conditional on this signal function are equivalent and so there is no loss of efficiency. On the other hand, the conditional likelihood is just as intractable and so there is nothing lost and nothing gained.

[12]The intractability of our log likelihood function seems inevitable in our context. That is, we do not believe there is some reasonable alternative model which could be used to estimate demand in our context whose estimation could be accomplished via maximizing a conditional log likelihood function. The reason is that we want to allow consumers to be able to purchase arbitrary bundles. When allowing bundling and even making the restrictive assumption that at most one unit of each good can be purchased one is left with 2^L possible bundles which can be purchased. It seems unlikely that there is an approach for calculating the probability of purchasing one of these 2^L bundles without explicitly calculating the utility of all the bundles. But, for a full-scale implementation of our method (where L can be over 1,000) this is a non-starter. To be overly dramatic, there are fewer than 2^{84} atoms in the observable universe. Thus we're dealing with numbers which are more than "astronomically" large.

At the other extreme we can specify $S(\boldsymbol{q}) = \{\boldsymbol{q}\}$ for all $\boldsymbol{q}$. Now, the probability mass function of $\boldsymbol{c}(\boldsymbol{p})$ conditional on $S(\boldsymbol{c}(\boldsymbol{p}))$ is particularly tractable. In fact, it is just

$$g(\boldsymbol{q}|\boldsymbol{q}') = \begin{cases} 1, & \text{if } \boldsymbol{q} = \boldsymbol{q}' \\ 0, & \text{else} \end{cases}$$

This function is easy to calculate but useless for estimating the parameters of our model (the parameters don't even enter the function).

The point of these extreme examples is to emphasize the trade-off inherent in designing S. If S passes too much information one losses efficiency (even to the point of losing consistency) and if S doesn't pass enough information than the log-likelihood function will remain intractable.

3.1 Signal Functions

We have just talked at length in intuitive terms about signal functions. Here we define the concept formally.

Definition 1. For each $\boldsymbol{q} \in \mathbb{N}_0^L$ let $S(\boldsymbol{q})$ be a random subset of $\mathbb{N}_0^L$. S is a *signal function* if (i) each realization of $S(\boldsymbol{q})$ contains $\boldsymbol{q}$, (ii) for each $Q \subseteq \mathbb{N}_0^L$ and all $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q$,

$$P\Big(S(\boldsymbol{q}) = Q\Big) = P\Big(S(\tilde{\boldsymbol{q}}) = Q\Big) \tag{14}$$

and (iii) for each $\boldsymbol{q} \in \mathbb{N}_0^L$ there is a countable collection $\tilde{\mathcal{Q}}$ consisting of subsets of $\mathbb{N}_0^L$ where $P\big(S(\boldsymbol{q}) \in \tilde{\mathcal{Q}}\big) = 1$.

Let $\boldsymbol{c}(\boldsymbol{p})$ be a random consumption bundle with a probability mass $f(\boldsymbol{q}|\boldsymbol{p})$ given by (7), let S be a signal function, and let $\mathcal{S} \equiv S(\boldsymbol{c}(\boldsymbol{p}))$. Recall that, intuitively speaking, $\mathcal{S}$ is providing a hint as to the true $\boldsymbol{c}(\boldsymbol{p})$ which was passed to it. That is, $\mathcal{S}$ gives a list of bundles where one of the bundles is the true $\boldsymbol{c}(\boldsymbol{p})$ bundle while the others are "impostors". Item (i) in the definition of the signal function just requires that $\mathcal{S}$ include the true bundle among the impostors. Item (ii) is a condition which ensures that the probability mass function of $\boldsymbol{c}(\boldsymbol{p})$ conditional on $\mathcal{S}$ is actually the same as the probability mass function of the consumer who is required to select a bundle form the choice set $\mathcal{S}$. In other words, item (ii) is what ensures that conditioning on the signal function actually produces tractable probability mass functions. Finally, item (iii) in the definition of the signal function is a technical condition which ensures $\mathcal{S}$ is measurable.

We have claimed that signal functions allow us to make the conditional probability mass function of demand more tractable. The following proposition will make this claim concrete. First, we require a definition. For a signal function S the support of S, denoted $\mathcal{Q}$, is defined by $\mathcal{Q} = \{Q \subseteq \mathbb{N}_0^L : \exists \boldsymbol{q} \in \mathbb{N}_0^L \text{ so that } P(Q \in S(\boldsymbol{q})) > 0\}$.

Proposition 6. *Let $\boldsymbol{c}$ be a random vector on $\mathbb{N}_0^L$ and $\boldsymbol{\rho}$ be random vector on $\mathbb{R}_{++}^L$. Let S be a signal function with support $\mathcal{Q}$ and assume $S(\boldsymbol{c})$ and $\boldsymbol{\rho}$ are independent conditional on $\boldsymbol{c}$. Let $Q \subseteq \mathbb{N}_0^L$ and let $f(\boldsymbol{q}|Q, \boldsymbol{p})$ denote the probability that $\boldsymbol{c} = \boldsymbol{q}$ conditional on $S(\boldsymbol{c}) = Q$ and $\boldsymbol{\rho} = \boldsymbol{p}$. Now,*

$$f(\boldsymbol{q}|Q, \boldsymbol{p}) = P\Big(\boldsymbol{c} = \boldsymbol{q}\Big|\boldsymbol{c} \in Q, \boldsymbol{\rho} = \boldsymbol{p}\Big), \qquad \text{for all } Q \in \mathcal{Q} \tag{15}$$

Consequently, if the probability mass function of $\boldsymbol{c}$ conditional on $\boldsymbol{\rho}$, denoted $f(\boldsymbol{q}|\boldsymbol{p})$, satisfies (7)*, then, for all $Q \in \mathcal{Q}$,*

$$f(\boldsymbol{q}|Q, \boldsymbol{p}) = \begin{cases} \dfrac{\exp(U(\boldsymbol{q}, \boldsymbol{p}))}{\sum_{\tilde{\boldsymbol{q}} \in Q} \exp(U(\tilde{\boldsymbol{q}}, \boldsymbol{p}))}, & \text{for } \boldsymbol{q} \in Q \\ 0, & \text{else.} \end{cases} \tag{16}$$

The punchline of Proposition 6 is that the denominator in (16) is much smaller than the denominator in (7). Thus, while we must give up hope of calculating likelihoods based on (7) we can indeed tractably calculate likelihoods using signal functions.

In Section A.1 of the Appendix we provide two conditions on the signal function which help ensure that conditional maximum likelihood estimation is consistent. A signal function which satisfies these two conditions is said to be *small and distinguishing*. Our consistency result, which we now go on to discuss, relies on S satisfying these two properties.

3.2 Consistent Estimation

In this subsection we present a consistency theorem for our estimator. We shall introduce assumptions on the data and on the signal function used which guarantee the consistency of our conditional maximum likelihood estimator. One point to bear in mind is that we are presenting our results as they apply to time series data on a single consumer. In our implementation we in fact have panel data and allow for individual-specific values of $\boldsymbol{b}$, d_1, d_2, and d_3. Our consistency result will still apply in this case provided we assume that the number of observations tends to infinity while the number of consumers stays fixed.

We now provide notation and assumptions which allow us to present our consistency theorem.

Suppose we have some data on N consumption bundles purchased by a single consumer and the prices of the products $(\boldsymbol{c}_1, \boldsymbol{\rho}_1), (\boldsymbol{c}_2, \boldsymbol{\rho}_2), \ldots, (\boldsymbol{c}_N, \boldsymbol{\rho}_N)$. Let $I \in \mathbb{N}$. We also assume that we have the ability to generate NI signal functions

$$S_{1,1}, S_{1,2}, \ldots, S_{1,I}, S_{2,1}, S_{2,2}, \ldots, S_{N,I}$$

For each n and i let $\mathcal{S}_{n,i} = S_{n,i}(\boldsymbol{c}_n)$. We shall refer to $\mathcal{O}_n = [\boldsymbol{c}_n, \boldsymbol{\rho}_n, \mathcal{S}_{n,1}, \mathcal{S}_{n,2}, \ldots, \mathcal{S}_{n,I}]$ as observation n. We estimate $\boldsymbol{\theta}$ by maximizing the following log likelihood function.

$$\mathcal{L}_N(\boldsymbol{\theta}) = \frac{1}{NI} \sum_{n=1}^{N} \sum_{i=1}^{I} \ln \Big(f(\boldsymbol{c}_n | \mathcal{S}_{i,n}, \boldsymbol{\rho}_n; \boldsymbol{\theta}) \Big) \tag{17}$$

where $f(\boldsymbol{c}_n | \mathcal{S}_{i,n}, \boldsymbol{\rho}_n; \boldsymbol{\theta})$ is defined by (16). Let $\hat{\boldsymbol{\theta}}_N \in \Theta$ satisfy

$$\hat{\boldsymbol{\theta}}_N \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_N(\boldsymbol{\theta}) \tag{18}$$

when this argmax exists. From Proposition 3 we know that $\psi(\boldsymbol{\theta})$ is the only part of $\boldsymbol{\theta}$ which we can hope to learn about. We provide assumptions under which $\psi(\hat{\boldsymbol{\theta}}_N)$ consistently estimates $\psi(\boldsymbol{\theta})$.

Assumption 1. The observations $\mathcal{O}_1, \mathcal{O}_2, \ldots$ are independent and identically distributed. Further, for each n,

$$\boldsymbol{\rho}_n, \ \mathcal{S}_{n,1}, \ \mathcal{S}_{n,2}, \ \ldots, \ \mathcal{S}_{n,I}$$

are independent conditional on $\boldsymbol{c}_n$.

Assumption 2. There is a non-empty open set $\mathcal{P} \subseteq \mathbb{R}^L_{++}$ so that for any non-empty open set $\mathcal{P}' \subseteq \mathcal{P}$ we have $P(\boldsymbol{\rho}_n \in \mathcal{P}') > 0$ for all n.

Assumption 3. The signal function $S_{n,i}$ is small and distinguishing for all n and i.

Assumption 4. There exists parameters $\boldsymbol{\theta}^* = [\boldsymbol{A}^*, \boldsymbol{b}^*, d_1^*, d_2^*, d_3^*] \in \Theta$ so that, for each n, the consumption bundle $\boldsymbol{c}_n$ has conditional probability mass function $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}^*)$ defined by (7). That is, $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}^*) = P(\boldsymbol{c}_n = \boldsymbol{q} | \boldsymbol{\rho}_n = \boldsymbol{p})$. Further, $\boldsymbol{A}^*$ has rank K.

Assumption 5. For each n, the following moments exist and are finite

$$\boldsymbol{E}[\boldsymbol{c}_n \boldsymbol{c}_n'], \qquad \boldsymbol{E}[\boldsymbol{\rho}_n \boldsymbol{\rho}_n'], \qquad \boldsymbol{E}[\boldsymbol{c}_n' \boldsymbol{c}_n \boldsymbol{\rho}_n], \qquad \text{and} \qquad \boldsymbol{E}[\boldsymbol{c}_n' \boldsymbol{c}_n \boldsymbol{\rho}_n' \boldsymbol{\rho}_n]$$

Assumption 1 puts independence restrictions on the data and signal functions. Assumption 2 is about the support of the price vectors. Assumption 3 requires that the signal functions used are small and distinguishing (see Section A.1 of the Appendix). Assumption 4 requires that the consumption bundles are generated by a random Qua consumer. Finally, Assumption 5 requires certain moments exist. We now state our consistency result.

Theorem 1. *Suppose assumptions 1-5 hold. Let* $\hat{\boldsymbol{\theta}}_N$ *satisfy* (18) *when the argmax is non-empty. Let* ψ *be defined by* (8)*. Then,*

$$\psi(\hat{\boldsymbol{\theta}}_N) \xrightarrow{a.s.} \psi(\boldsymbol{\theta}^*) \tag{19}$$

Two remarks are in order. First, by Proposition 3, we know that $\psi(\boldsymbol{\theta})$ is all that we can hope to learn about $\boldsymbol{\theta}$. Thus, Theorem 1 shows that our estimator recovers (almost surely) the features of interest of $\boldsymbol{\theta}$. Second, while the proof of Theorem 1 is inspired by the proof of Theorem 2.7 (on consistency without compactness) in Newey and McFadden (1994), there are two features of our problem which prevent a straightforward application this result. These features are (i) the parameters in $\boldsymbol{\theta}$ are not identified (only $\psi(\boldsymbol{\theta})$ is identified) and (ii) the objective function in (17) is not concave.[13] Thus, our proof has to adapt the proof in Newey and McFadden (1994) to handle these difficulties.

4 Algorithm design and simulation

Theorem 1 describes a consistent estimator for the parameters of the model. To implement this, we will need to minimize $-\mathcal{L}_N(\boldsymbol{\theta})$, which is defined in (17). In this Section we define an algorithm to do this, and observe its performance on simulated data.

4.1 Computational Considerations

Gradient Descent would be an appropriate numerical algorithm for us, if it were not for computation constraints. To get around this, we adapt Gradient Descent, making it suitable for our machines. In particular, at each iteration in the Gradient Descent we

[13]The objective function would be concave if the utility function U was linear in parameters. In fact, the transformed parameters $\psi(\boldsymbol{\theta})$ do enter U linearly and so one might think that this problem can be solved by transforming the parameter space. The problem is that the image $\psi(\Theta)$ is not convex and so again Theorem 2.7 would not directly apply.

approximate the gradient of (17) with respect to the parameters, rather than computing it exactly, by averaging across only a subset of the observations in our dataset, and not across all N observations. In the machine learning literature, such a subset is known as a *batch*. To exploit processors efficiently we set batch size equal to a power of two: in practice, to $2^{10} = 1,024$ observations. Typically, the batch may contain under 1% of the dataset.

At each step in our Gradient Descent we access a new batch, and for each observation in the batch we take a new draw from a signal function. We will call such a realization of a signal function on a datapoint a *signal set*. We use the small and distinguishing signal function detailed in the Appendix, configured to produce signal sets of size 100 and quantities of goods no greater than six.

This approach is substantially inspired by the method of negative sampling for word embeddings outlined in Mikolov et al. (2013). Following that literature, we do not draw batches independently with replacement from our dataset, but instead we just access batches sequentially as we run through the dataset. Finally, we record the obtained gradient and use it in immediately subsequent steps according to the `adam` algorithm of Kingma and Ba (2015).

When the whole dataset has been accessed in this way once, we say that an *epoch* is completed, and we begin again to pass through our data, generating fresh signal sets for each revisited batch.

In advance of fitting, we seed all our unconstrained real parameters with independent random normals of variance $1/K$, and all positive parameters with exponentially-distributed random variables of mean $1/K$.

All this is specified in the code made public at `www.github.com/jeremy-large/RUBE`.

4.2 Simulation Design

Our algorithm may be described as a form of Stochastic Gradient Descent. Because batch size is limited by processing power, it is appropriate to do simulations to understand the algorithm's small-sample properties. This requires us to simulate from a postulated model, M, then to fit using this simulated data, and finally to confirm that we rediscover the parameters of M.

To this end, we set M to a model previously fitted to our scanner dataset detailed in Section 5, whose parameters we call θ^*. We then simulate bundles drawn from M. This involves some assumptions:

- we hold prices constant at their empirical mean in our data;
- we study a single consumer/user;
- we set to zero the period-specific dummy terms presented in Section 2.6;
- we limit our attention to bundles containing exactly nine separate goods.

Under these assumptions not all parameters of M are identified: for example, in the absence of price fluctuations we cannot discern d_3. Nevertheless, we apply a Metropolis-Hastings algorithm defined as follows:

1. Draw an initial bundle, $\boldsymbol{q}_0$, uniformly from $[0, 1, 2, 3, 4, 5, 6]^L$ subject to having exactly nine non-zero entries. Set $\boldsymbol{q} = \boldsymbol{q}_0$.
2. Use the small and distinguishing signal function detailed in the Appendix to draw a minimally small signal set, $S(\boldsymbol{q})$, of size only 2. So, $S(\boldsymbol{q}) = \{\boldsymbol{q}, \tilde{\boldsymbol{q}}\}$ and contains both $\boldsymbol{q}$, and one proposed *negative sample*, $\tilde{\boldsymbol{q}}$.

   Definition 1 implies directly that the proposal distribution, $\tilde{\boldsymbol{q}}|\boldsymbol{q}$, is symmetric in $\tilde{\boldsymbol{q}}$ and $\boldsymbol{q}$; furthermore this signal function preserves the number of distinct items in the bundle.
3. Evaluate the utility of $\tilde{\boldsymbol{q}}$, and calculate its difference to the utility of $\boldsymbol{q}$. The Metropolis-Hastings acceptance ratio is the exponent of this difference.[14]
4. If we accept $\tilde{\boldsymbol{q}}$ we set $\boldsymbol{q} = \tilde{\boldsymbol{q}}$, otherwise $\boldsymbol{q}$ is unchanged.
5. Return to 2.

To give better exposure to the ergodic distribution, we wait for 2,500 steps before sampling from this process at point 4. In order to limit autocorrelation, we subsequently sample only every 50 steps.

[14]This follows because utilities are logits in our model.

4.3 Simulation Results

We run two simulations:

- a Precise Simulation, with larger than normal batch size of 8,192 and smaller than normal universe of $L = 50$. Here we expect to see good performance.
- a Commensurate Simulation with a standard batch size of 1,024 and a vocabulary size of $L = 2,500$. Here we expect to learn about performance at scales comparable to that of our empirical implementation.

Theorem 1 states that when the function $\psi(\cdot)$ is applied to our fitted parameter values, we have convergence almost surely to the true parameters. After epoch i, let us call our fitted parameters $\hat{\boldsymbol{\theta}}^i$. Then we will be interested in the timeseries of $\{\psi(\hat{\boldsymbol{\theta}}^i) : i = 1, 2, ...\}$. We expect this to approach the truth, $\psi(\theta^*)$. However, the incessant addition of fresh randomness in every epoch, i, will introduce ongoing ergodic perturbations around the truth.

The function $\psi(\cdot)$ is defined in (8). Its first term is a symmetric matrix containing pairwise inner products of the goods' attribute vectors. The second term contains the pairwise inner products of the customer's, $\boldsymbol{b}$ (their 'embedding') with the attribute vectors ('embeddings') of the goods. The third and fourth terms are scalars, d_1 and d_2, governing the dis-utility of expenditure. We study the timeseries of these items across epochs.

4.3.1 Towards interpreting our simulation results

We began by choosing a model, M. By using the parameters of M, we simulated some data to test our algorithm. To make this exercise as relevant as possible, the parameters of M were actually ones fitted to the DunnHumby scanner data, which we describe in detail in Section 5. To assist in interpretation, we order the DunnHumby goods in decreasing prevalence; and we pick the most frequent of these goods as the ones to appear in our simulations. Table 2 describes the top 25 goods.

In studying the two simulations, we will pay particular interest to $(\boldsymbol{A}'\boldsymbol{A})_{1,1}$, which, because it is $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1$, is the squared magnitude of the attribute vector for the most common product, a milk. We also observe $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2$ which is the inner product of the vectors of the top two products here. Finally, $\boldsymbol{\alpha}_4'\boldsymbol{\alpha}_5$ will be of interest, because it is negative: loosely, following the discussion in Section 2, we may say that shredded cheese and large eggs are

	Product	Units	Manufacturer code
1	FLUID MILK WHITE ONLY	GA	69
2	BANANAS	LB	2
3	FLUID MILK WHITE ONLY	n/a	69
4	SHREDDED CHEESE	OZ	69
5	EGGS - X-LARGE	DZ	69
6	MAINSTREAM WHITE BREAD	OZ	69
7	SOFT DRINKS 12/18&15PK CAN CAR	OZ	1208
8	POTATO CHIPS	OZ	544
9	SOFT DRINKS 12/18&15PK CAN CAR	OZ	103
10	SFT DRNK 2 LITER BTL CARB INCL	LTR	103
11	HAMBURGER BUNS	OZ	69
12	MAINSTREAM WHITE BREAD	OZ	910
13	SFT DRNK 2 LITER BTL CARB INCL	LTR	69
14	SALAD BAR FRESH FRUIT	n/a	2
15	STRAWBERRIES	OZ	5937
16	SFT DRNK 2 LITER BTL CARB INCL	LTR	1208
17	CANDY BARS (SINGLES)(INCLUDING	OZ	693
18	FRZN BAGGED VEGETABLES - PLAIN	OZ	69
19	SFT DRNK SNGL SRV BTL CARB (EX	OZ	103
20	BEERALEMALT LIQUORS	OZ	239
21	HOT DOG BUNS	OZ	69
22	CIGARETTES	PK	539
23	ONIONS SWEET (BULK&BAG)	LB	2
24	GRAPES RED	LB	2
25	HEAD LETTUCE	CT	673

Table 2: The 25 most frequent goods in the DunnHumby dataset after cleaning, in decreasing order of frequency. The Manufacturer Code is a unique anonymous identifier of the manufacturer. In simulating our model, we will pay particular interest to $(\boldsymbol{A}'\boldsymbol{A})_{1,1}$, which, because it is $\boldsymbol{\alpha_1}'\boldsymbol{\alpha_1}$, is the magnitude of the embedding vector for the most common good, at the top of this list, a milk. We also study $\boldsymbol{\alpha_1}'\boldsymbol{\alpha_2}$ which is the inner product of the vectors of the top two goods here (milk/bananas). Finally, $\boldsymbol{\alpha_4}'\boldsymbol{\alpha_5}$ will be of interest: it relates purchases of shredded cheese and of extra-large eggs.

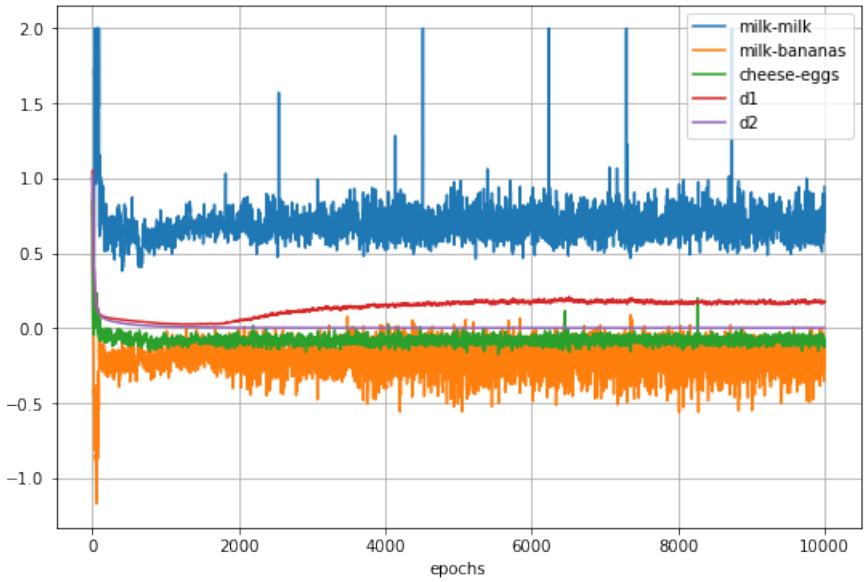

Figure 1: Precise Simulation: the evolution of various estimates as epochs pass during the estimation. Parameter estimates greater than 2 are replaced with 2. The true values are as follows: $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0.83$ ('milk-milk') ; $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = -0.15$ ('milk-bananas'); $\boldsymbol{\alpha}_4'\boldsymbol{\alpha}_5 = -0.05$ ('cheese-eggs'); $d_1 = 0.28$; $d_2 = 0.0002$.

co-present in bundles, perhaps related to their use together in standard cooking recipes such as for omelets.

4.3.2 Precise simulation: results

We collected 41,700 simulated bundles, each of which contained nine items from among a universe of 50 goods. It was taken as known that $K = 16$. We then ran the fitting algorithm for 10,000 epochs. There were 819 parameters to estimate.

Figure 1 plots certain parameter values as epochs pass during estimation. Estimates establish ergodic distributions after around 4,000 epochs.

To assess how close these distributions are to the truth, we average fitted parameter values in the final 100 epochs, and plot these against true parameter values in Figures 2 and 3. Over all, and pending better methods of inference, the fit seems quite acceptable in this Precise Simulation.

4.3.3 Commensurate simulation: results

We now turn to a more challenging simulation, which is more in line with our empirical objectives. This time, 2,500 goods are present: we must estimate attribute vectors for

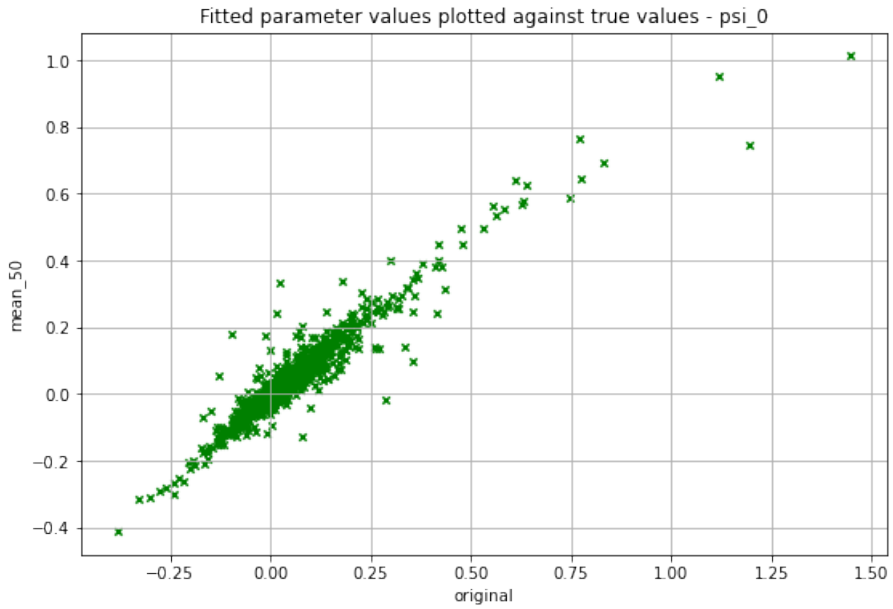

Figure 2: Precise Simulation: a cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true (original) values. We average fitted parameter values in the final 100 epochs, and plot these against true parameter values. The 1,275 quantities plotted here are $\{\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j : 1 \leq i \leq j \leq 50\}$.

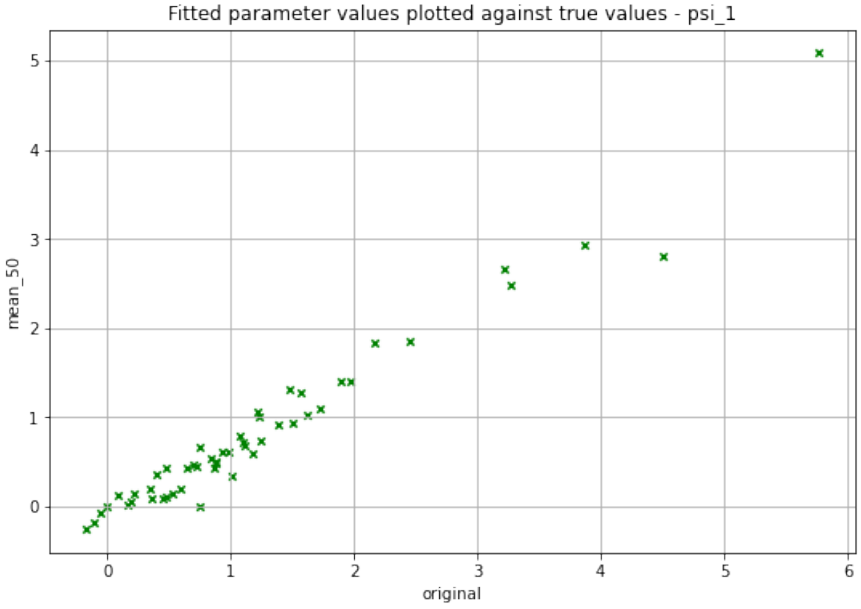

Figure 3: Precise Simulation: a cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true (original) values. We average fitted values in the final 100 epochs, and plot these against true parameter values. The 50 quantities plotted here are $\{b'\boldsymbol{\alpha}_j : 1 \leq j \leq 50\}$.

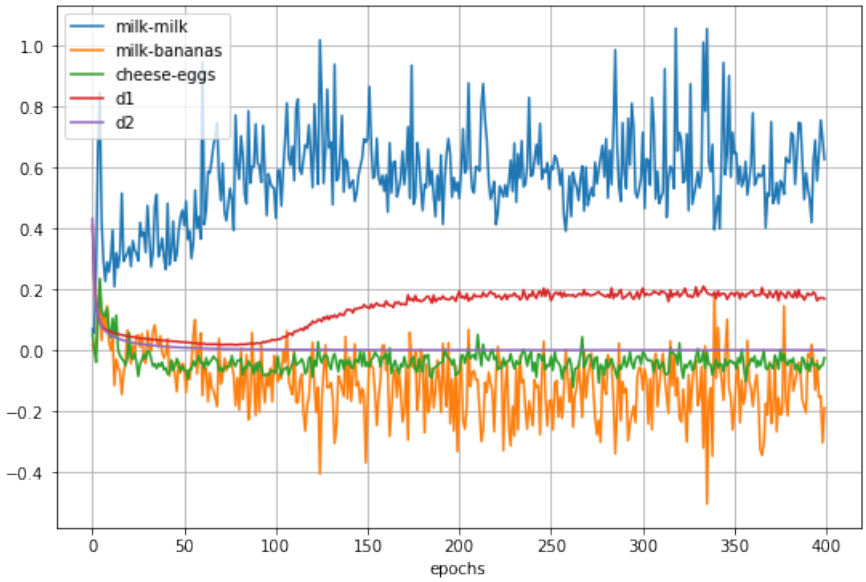

Figure 4: Commensurate Simulation: the evolution of various estimates as epochs pass during the estimation. The true values are as follows: $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0.83$ ('milk-milk') ; $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = -0.15$ ('milk-bananas'); $\boldsymbol{\alpha}_4'\boldsymbol{\alpha}_5 = -0.05$ ('cheese-eggs'); $d_1 = 0.28$; $d_2 = 0.0002$.

every one of these. We again take it as known that $K = 16$, so that in all there were 40,019 parameters to estimate. 112,000 simulated bundles were collected, which is 16% more than we will have access to in our empirical application. Each bundle again contained nine items (in various quantities). We report on the first 400 epochs.

Figure 4 plots some parameter values as epochs pass during estimation. This is a less stable picture than Figure 1, which is directly comparable. But by eye, it seems that after around 200 epochs estimates are settling into ergodic distributions. This is noticeably sooner than in the Precise Simulation, due presumably to the larger dataset and perhaps the greater variety of bundles.

As we did for the Precise Simulation we average fitted parameter values over the final 100 epochs, and plot these against true parameter values in Figures 5 and 6. To be directly comparable, these figures report information only for the set of goods which also appear for the Precise Simulation in Figures 2 and 3 (although there is a vast set of additional parameter values which were estimated at the same time). Over all, and pending better methods of inference, the fit is not as good as in the Precise Simulation despite our larger dataset, but is acceptable.

We will see in Section 5 that our real dataset will be about the same size as the one

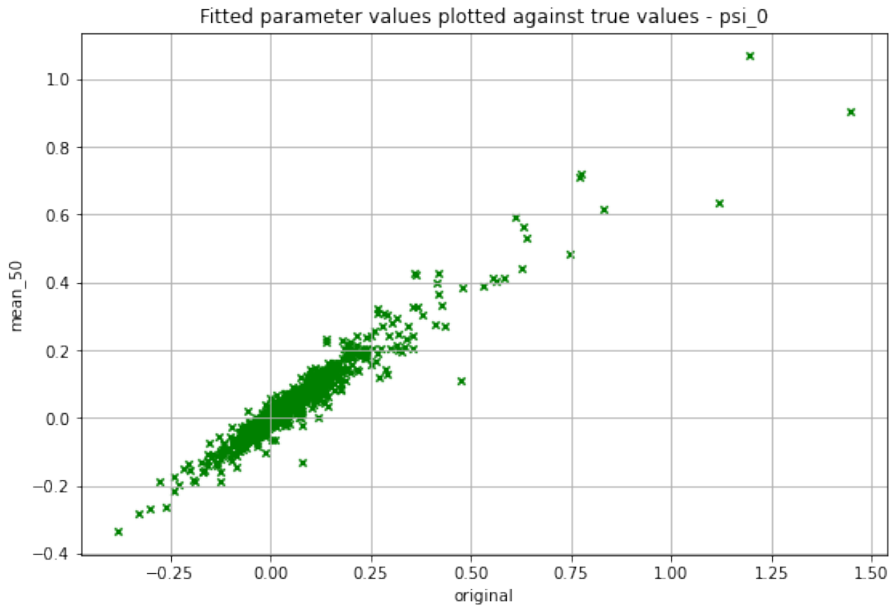

Figure 5: Commensurate Simulation: a cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true (original) values. We average fitted values in the final 100 epochs, and plot these against true parameter values. The 1,275 quantities plotted here are $\{\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j : 1 \leq i \leq j \leq 50\}$.

used in this simulation and that we will use early stopping criteria to aid in judging a suitable number of epochs to use in estimation. However, given this simulation's smaller dataset, fitting for 400 epochs on the real dataset would appear appropriate, although perhaps a little on the high side.

5 Empirical application

We apply our model to scanner data made available by DunnHumby, a customer data science company. The dataset, named 'The Complete Journey' by DunnHumby, details approximately 2.6m purchases by 2,500 households in the US over a two-year period. The households, described as frequent shoppers at a retailer, were recruited to a program whereby they recorded their purchases.

The start and end dates of this dataset are not made available to us, but days and weeks are clearly identified by number. We exclude from our study the final 26 weeks of the data, for a future out-of-sample validation exercise. We have never considered this data. We also exclude the first 15 weeks, during which time new households continued to join the program. We study the remaining 1.6m purchases, and find that they were

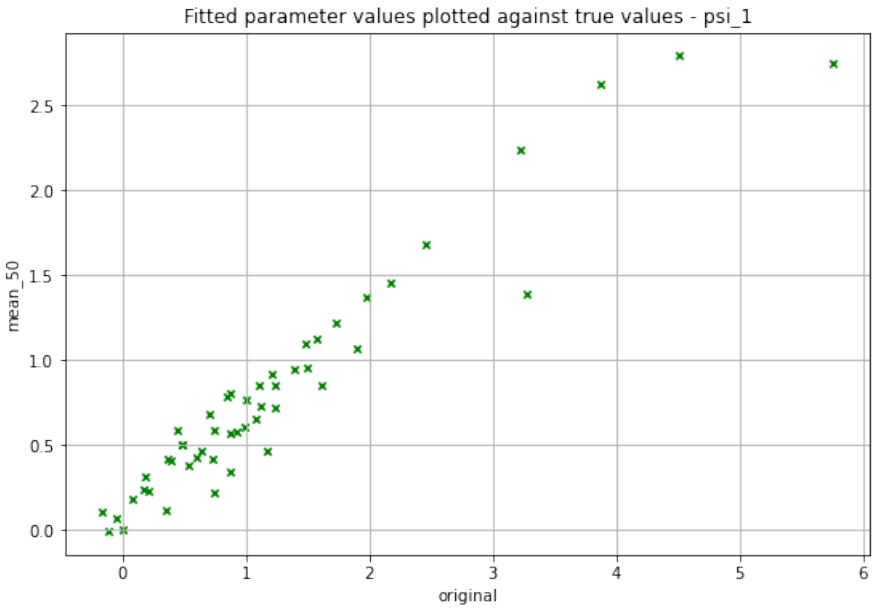

Figure 6: Commensurate Simulation: a cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true (original) values. We average fitted values in the final 100 epochs, and plot these against true parameter values. Although there are estimates for 2,500 goods, the 50 quantities plotted here are $\{\boldsymbol{b}'\boldsymbol{\alpha}_j : 1 \leq j \leq 50\}$.

made on 178,212 separate check-out occasions over 60 weeks.

We view each such check-out at the till as recording a single consumption bundle. On average, bundles cost $28.56 and contained 8 items. The most expensive bundle cost $961.49; and the biggest bundle contained 141 items.

For each good, we observe an SKU code, store sub-department information, and anonymous manufacturer and brand codes. We also observe a product description such as for example `CONDIMENTS/SAUCES|STEAK & WORCHESTER SAUCE`, and in some cases a product size in various units such as ounces or gallons. We consider each household to be a single consumer.

When we distinguish products by their descriptions, manufacturers, brands and departments, as well as by their sizes bucketed into 20% bins, we find 24,680 goods. On average, each of these goods is mentioned 60 times in our data.

We remove a long tail of 19,777 goods which are purchased fewer than 50 times. This done, we remove the 200 low-paying consumers who still spent on average less than $10 per visit. We finally remove a further 1,440 consumers, who went to the retailer fewer than 50 times. After this procedure, we are left with a vocabulary of 4,737 goods and

986 consumers who formed from these goods 96,285 consumption bundles.

We estimate individual-specific preference parameters, $[b_u, d_{1u}, d_{2u}, d_{3u}]$, for these consumers, whom we index throughout this Section with $\{u \in \mathbb{N}_+ : u \leq 986\}$. DunnHumby provides self-reported demographic household information for 588 of these individuals.

The model we fit includes the period-specific dummy terms introduced in Section 2.6. We break our dataset into four-week periods, of which there are 15, and for the jth such period, τ_j, we therefore estimate a period-specific term, $\boldsymbol{b}_{\tau_j}$, which is a vector in $\mathbb{R}^K$ describing temporary surges in any of the K attributes' desirability, that are common to all consumers in the four-week period. We impose the constraint that this term's average value across our 15 periods is $\mathbf{0} \in \mathbb{R}^{\mathbf{K}}$.

For each consumption bundle in our data, we know the period, τ_j, in which it occurred. In our fitting procedure we repeatedly apply the signal function described in Section 4 to the bundle. The resulting signal sets contain false bundles which include other goods, that were not in the true bundle. We set the price of these other goods to the average price at which they were purchased during period τ_j.[15]

Altogether, the model has 94,766 parameters. We use the code at `www.github.com/jeremy-large/RUBE` to estimate it. This uses a stochastic gradient descent method. Section 4 specifies and simulates this arrangement in detail.

5.1 Validation accuracy

We hold-out a small sliver, two per cent, of our data for validation testing. We equip each bundle in this validation dataset with two fixed signal sets. During estimation, we consider an ancillary classification exercise where after each epoch, we assess validation accuracy. Thus, for each signal set in the validation dataset, we identify the bundle with the greatest estimated utility; and we count how often this is the true bundle. This may be thought of as Bayes classification using interim, plug-in, parameter values.

When all parameters of the model are set to zero, this classification exercise yields a validation accuracy of 1%. With randomly-seeded parameters as detailed in section 4.1, this accuracy stands at 7%. But, after around 150 epochs, validation accuracy plateaus at around 30%. About a third of the time, therefore, our classifier is able to identify the truth from among a set of 100 possible bundles, which were not used to fit it. We stop

[15] In the event that the other good was not at all purchased during τ_j, we set its price to its average across our full dataset.

self-reported household income p.a. ($)	count	mean	std	25%	50% median	75%
175K+	21	0.21	0.12	0.15	0.2	0.27
150-174K	23	0.22	0.09	0.16	0.23	0.28
125-149K	28	0.29	0.18	0.15	0.22	0.4
100-124K	26	0.33	0.11	0.23	0.32	0.41
75-99K	66	0.29	0.15	0.2	0.25	0.38
50-74K	140	0.35	0.17	0.25	0.32	0.41
35-49K	118	0.38	0.19	0.24	0.35	0.49
25-34K	58	0.42	0.22	0.27	0.4	0.54
15-24K	57	0.43	0.21	0.27	0.36	0.58
Under 15K	51	0.39	0.16	0.26	0.38	0.5

Table 3: Distribution of d_1 across households reporting demographic information. We estimate household-specific preference parameters, $[b_u, d_{1u}, d_{2u}, d_{3u}]$, for 986 households, u. DunnHumby provides self-reported demographic information for 588 of these households, which partitions them by income band. The table displays income bands by decreasing self-reported affluence. Means, standard deviations, and quantiles of d_{1u} (the linear coefficient on expenditure) are shown and are broadly increasing as affluence falls. The categories, `250K+, 200-249K, 175-199K` contain small numbers of observations and have been coalesced.

the algorithm after 400 epochs.

5.2 Estimated parameters and results

As we see, then, estimating the 94,766 parameters of our empirical model materially improves forecast accuracy on a reserved dataset. But it is also possible to gain insight by interpreting these estimated parameters. In this section, we report on several estimates selected for their illustrative interest, namely:

1. The distribution of $\widehat{\boldsymbol{d_1}}$, which estimates linear dis-utility of expenditure for various households.

2. Properties of $\widehat{\boldsymbol{\alpha_{21}}}$, the attribute vector of Good 21, as described at the base of Table 2. This good is the most commonly purchased **hot dog bun** in our dataset. It is interesting to look at Good 21, because of our informal knowledge that it is used in conjunction with other goods, such as hot dogs, sausages, salsa, and even burgers.

3. Large elements of the vector $\widehat{\boldsymbol{A}}_{\mathbf{1}}$ (which is the first row of $\widehat{\boldsymbol{A}}$). This vector evaluates the first attribute, which has a special role capturing product-specific dis-utility of expenditure. If $\widehat{A}_{1\ell}$ is high for some good ℓ, then consumers, u, with $d_{3u} > 0$ are particularly disinclined to incur expense when buying it.

5.2.1 Linear dis-utility of expenditure

Table 3 compares our estimates of d_{1u}, the linear term capturing the dis-utility of expenditure, with self-reported household income for the 588 households where these overlap. We see that the mean estimate of d_{1u} is 0.21 for those households, u, reporting annual earnings in excess of \$175k, but that this rises to 0.39 for those reporting annual earnings below \$15k. As we might have expected, therefore, those households who say that their income is high are less deterred from consumption by its expense.

5.2.2 Demand for hot dog buns

Among the most frequently purchased items listed in Table 2, the 21st item, a hot dog bun, yields interesting substitution patterns with other goods. Indeed, its description alone indicates that it is normally thought to be used in concert with at least one other good, namely a hot dog. We will call it *Good 21* below.

We study Good 21 in two ways, both of which calculate *cosine similarities*. Let us define the cosine similarity between good ℓ and good j as the cosine of the angle between vectors $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}}$ and $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{j}}$. It is therefore an estimate, which lies in the interval $[-1, 1]$.

A cosine similarity of 1 indicates that two vectors are identical up to a positive constant, while a similarity of -1 indicates vectors which point in opposite directions. Because it is a normalization of the dot products $\boldsymbol{\alpha}_\ell' \boldsymbol{\alpha}_j$ discussed in Section 2, a large negative cosine similarity between two goods is likely to indicate it is normally desirable to purchase them together, if at all.

Table 4 exhibits the 35 goods in the data, which contain 'HOT DOG' in their description. The table is ordered by decreasing estimated cosine similarity to Good 21. The most similar tabulated goods to Good 21 are exactly the other 8 varieties of hot dog bun offered in store. Their estimated cosine similarities to it all exceed 0.3. The rest of the goods appearing in the list have lower cosine similarity to Good 21. They are many varieties of hot dogs, and chili sauce for hot dogs.

	Cosine Similarity	Manu-facturer	SUB-DESC	COMMODITY-DESC
0	1.000	69	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
10	0.834	69	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
12	0.812	69	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
19	0.793	69	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
25	0.784	910	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
35	0.744	910	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
199	0.565	1638	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
266	0.521	1838	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
802	0.331	69	HOT DOG BUNS	BAKED BREAD/BUNS/ROLLS
1006	0.275	93	KOSHER/SPECIALTY	HOT DOGS
3065	-0.159	69	HOT DOG CHILI SAUCE	MEAT - SHELF STABLE
3133	-0.17	665	PREMIUM - MEAT	HOT DOGS
3281	-0.199	83	HOT DOG CHILI SAUCE	MEAT - SHELF STABLE
3671	-0.29	1323	PREMIUM - MEAT	HOT DOGS
3678	-0.291	69	PREMIUM - MEAT	HOT DOGS
3726	-0.303	1094	KOSHER/SPECIALTY	HOT DOGS
4034	-0.387	69	PREMIUM - BEEF	HOT DOGS
4056	-0.39	593	HOT DOG CHILI SAUCE	MEAT - SHELF STABLE
4101	-0.406	1089	BETTER FOR YOU	HOT DOGS
4143	-0.419	1089	KOSHER/SPECIALTY	HOT DOGS
4167	-0.427	2012	KOSHER/SPECIALTY	HOT DOGS
4183	-0.433	1094	PREMIUM - MEAT	HOT DOGS
4358	-0.491	665	KOSHER/SPECIALTY	HOT DOGS
4436	-0.52	1323	PREMIUM - MEAT	HOT DOGS
4478	-0.536	1094	PREMIUM - BEEF	HOT DOGS
4514	-0.555	1323	BETTER FOR YOU	HOT DOGS
4591	-0.604	69	ECONOMY - MEAT	HOT DOGS
4594	-0.606	1719	BETTER FOR YOU	HOT DOGS
4646	-0.646	1089	PREMIUM - BEEF	HOT DOGS
4668	-0.671	1089	PREMIUM - BEEF	HOT DOGS
4693	-0.715	5226	KOSHER/SPECIALTY	HOT DOGS
4725	-0.781	1323	PREMIUM - BEEF	HOT DOGS
4726	-0.783	1425	ECONOMY - MEAT	HOT DOGS
4732	-0.835	1089	PREMIUM - MEAT	HOT DOGS
4733	-0.837	1323	PREMIUM - MEAT	HOT DOGS

Table 4: The 35 goods of the 4,737 goods described in Section 5, which contain `'HOT DOG'` in their description. The table is ordered by decreasing estimated cosine similarity with Good 21, which is a prevalent make of hot dog bun. This is an interesting good to look at because it is likely to be used in conjunction with hot dogs. The most similar tabulated goods to Good 21 are exactly the other 8 varieties of hot dog bun offered in store. Their estimated cosine similarities to it all exceed 0.3. Of the goods in this list, which we might categorize intuitively as complements to hot dog buns, all but one appear with negative cosine similarity to them.

Token	Cosine	Manu-facturer	SUB-DESC	COMMODITY-DESC
3062	-0.783	69	HAMBURGER BUNS	BAKED BREAD/BUNS/ROLLS
4400	-0.789	1225	INSTANT BREAKFAST	CONVENIENT BRKFST/WHLSM SNACKS
1020	-0.799	2296	SHAMPOO	HAIR CARE PRODUCTS
2062	-0.807	3537	CHEESE: BULK	CHEESES
205	-0.815	2	SQUASH ZUCCHINI	SQUASH
118	-0.835	1089	PREMIUM - MEAT	HOT DOGS
1536	-0.837	1323	PREMIUM - MEAT	HOT DOGS
1262	-0.837	1094	PICKLES	MISCELLANEOUS
400	-0.84	1089	HAM	LUNCHMEAT
11	-0.865	69	HAMBURGER BUNS	BAKED BREAD/BUNS/ROLLS

Table 5: The 10 goods of the 4,737 goods described in Section 5 that have the most negative estimated cosine similarity to Good 21, which is a frequently purchased hot dog bun. The product-*Token* is a unique identifier of the good, and orders goods in roughly decreasing order of prevalence.

Table 5 describes the ten goods from among the 4,737 goods that have the most negative estimated cosine similarity to these hot dog buns, Good 21. It is not a surprise to see pickles, lunchmeat ham, sausages and hamburger-buns featured in this list.

The presence of shampoo in the list is more surprising: this may describe a coincidental co-occurrence in our data, or indeed a real propensity to co-purchase. It could be that $K = 16$ degrees of freedom is insufficient and causes bias here.

5.2.3 Product-specific dis-utility of expenditure

In (4), which describes the expected utility of a bundle, $\boldsymbol{q}$, part of the dis-utility of expenditure is accounted for by the term $-2d_3(\boldsymbol{A}_1'\boldsymbol{q})(\boldsymbol{p}'\boldsymbol{q})$. This is non-zero for any consumer, u, with $d_{3u} > 0$, who is considering a product, j with $A_{1j} > 0$. As such, $\widehat{\boldsymbol{A_1}}$ is a vector of estimates of product-specific dis-utilities of expenditure.

Table 6 details ten goods, ℓ, with high estimates $\widehat{A}_{1\ell}$. These tend to be durables such as paper towels or frozen fruit or pies, but can also be coupons which are redeemable for gasoline. Before stocking-up on e.g. paper towels perhaps consumers prefer to wait for a deal on the price.

Figures 7, 8, 9, 10, 11, and 12 contain further diagnostic information about the estimated fit. The Appendix contains some discussion of the choice $K = 16$.

product_token	$\hat{A}_{1\ell}$	Manufacturer	COMMODITY-DESC
3760	7.91	69	FROZEN FRUIT
3156	1.21	781	CROUTONS SALAD TOPPERS BREAD
3224	0.64	1208	CARBONATED WATER - FLVRD SWEET
1092	0.58	1276	NON-CARB FLVRD DRNKNG/MNRL WAT
2217	0.34	69	PAPER TOWELS & HOLDERS
672	0.31	754	PASTA SAUCE
441	0.27	69	GASOLINE COUPONS
1711	0.25	69	POULTRY LUNCHMEAT
687	0.25	69	PUDDING & GELATIN CUPS/CANS
398	0.24	69	FRZN WHIPPED TOPPING

Table 6: The ten goods, ℓ, with highest estimates $\widehat{A}_{1\ell}$. These tend to be durables such as bottled water, but can also be snacks. The parameter captures product-specific disutility of expenditure. Across the full dataset the upper 0.75 quantile of $\widehat{\boldsymbol{A}_1}$ is 0.007.

6 Concluding remarks

We have presented a random utility model and estimation technique suitable for analyzing large datasets. While our approach has many advantages there are also many areas requiring future work. We now outline those areas where we believe future work would be most beneficial.

First, we would like to know the asymptotic distribution of our estimator. Once this is known we could calculate standard errors, confidence intervals, and perform hypothesis tests. Second, more thought should be given to modeling unobserved preference heterogeneity. Currently, we take advantage of the fact that our dataset has a large number of observations per consumer to reduce concerns of an incidental parameters problem. We think future work should consider a version of our random utility model suitable for data where each consumer is not seen many times.

This work can be considered in the light of substantial recent advances in natural language processing, enabled by associating linguistic units such as words with vectors of attributes. It is motivated by an ambitious question: Can we look ahead to a time when, analogously, any attribute of a *good* that is significant for people can be identified and inferred from consumption data? We suggest that this is a reasonable aspiration.

References

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–85. https://doi.org/10.1257/aer.p20151021

Barnett, W. A., & Serletis, A. (2008). Consumer preferences and demand systems. *Journal of Econometrics*, *147*, 210–224.

Berry, S., & Haile, P. A. (2021). Foundations of demand estimation. In K. Ho, A. Hortaçsu, & A. Lizzeri (Eds.), *Handbook of industrial organization, volume 4* (pp. 1–62). Elsevier. https://doi.org/https://doi.org/10.1016/bs.hesind.2021.11.001

Berry, S., Khwaja, A., Kumar, V., Musalem, A., Wilbur, K., Allenby, G., Anand, B., Chintagunta, P., Hanemann, M., Jeziorski, P., & Mele, A. (2014). Structural models of complementary choices. *Marketing Letters*, *25*, 245–256. https://doi.org/10.1007/s11002-014-9309-y

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890. http://www.jstor.org/stable/2171802

Berry, S., Levinsohn, J., & Pakes, A. (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, *112*(1), 68–105. http://www.jstor.org/stable/10.1086/379939

Billingsley, P. (1986). *Probability and measure* (Second). John Wiley; Sons.

Blundell, R., Horowitz, J. L., & Parey, M. (2017). Nonparametric estimation of a nonseparable demand function under the Slutsky inequality restriction. *Review of Economic Studies*, *99*(2), 291–304.

Blundell, R., & Robin, J.-M. (2000). Latent separability: Grouping goods without weak separability. *Econometrica*, *68*(1), 53–84. http://www.jstor.org/stable/2999475

Chernozhukov, V., Hausman, J., & Newey, W. K. (2019). *Demand analysis with many prices* (CeMMAP working papers CWP59/19). Centre for Microdata Methods and Practice, Institute for Fiscal Studies. https://ideas.repec.org/p/ifs/cemmap/59-19.html

Deaton, A. S., & Muellbauer, J. N. (1980). An almost ideal demand system. *American Economic Review*, *70*, 312–326.

Deb, R., Kitamura, Y., Quah, J. K.-H., & Stoye, J. (Forthcoming). Revealed price preference: Theory and empirical analysis. *Review of Economic Studies*.

Dubois, P., Griffith, R., & O'Connell, M. (2020). How well targeted are soda taxes? *American Economic Review*, *110*(11), 3661–3704. https://doi.org/10.1257/aer.20171898

Gandhi, A., & Nevo, A. (2021). Empirical models of demand and supply in differentiated products industries. In K. Ho, A. Hortaçsu, & A. Lizzeri (Eds.), *Handbook of industrial organization, volume 4* (pp. 63–139). Elsevier. https://doi.org/https://doi.org/10.1016/bs.hesind.2021.11.002

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, *97*(3), 713–744. https://doi.org/10.1257/aer.97.3.713

Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models* (Vol. 1). Cambridge University Press. https://EconPapers.repec.org/RePEc:cup:cbooks:9780521477451

Greene, W. (2015). Panel data models for discrete choice. In B. H. Baltagi (Ed.), *The oxford handbook of panel data* (pp. 171–201). Oxford University Press.

Hausman, J. A., & Newey, W. K. (2016). Individual heterogeneity and average welfare. *Econometrica*, *84*(3), 1225–1248. https://doi.org/10.3982/ECTA11899

Hendel, I. (1999). Estimating multiple-discrete choice models: An application to computerization returns. *The Review of Economic Studies*, *66*(2), 423–446. http://www.jstor.org/stable/2566997

Keane, M. P. (2015). Panel data models for discrete choice. In B. H. Baltagi (Ed.), *The oxford handbook of panel data* (pp. 548–582). Oxford University Press.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations*. http://arxiv.org/abs/1412.6980

Kitamura, Y., & Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, *86*(6), 1883–1909. https://doi.org/10.3982/ECTA14478

Lewbel, A., & Pendakur, K. (2009). Tricks with Hicks: The EASI demand system. *American Economic Review*, *99*(3), 827–63. https://doi.org/10.1257/aer.99.3.827

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in frontiers in econometrics. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 2373–2375). Academic Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations*. http://arxiv.org/abs/1301.3781

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, *69*(2), 307–342. http://www.jstor.org/stable/2692234

Newey, W. K., & McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. Elsevier. https://doi.org/https://doi.org/10.1016/S1573-4412(05)80005-4

Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press.

Ruiz, F. J. R., Athey, S., & Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *Ann. Appl. Statist.*, *14*(1), 1–27. https://doi.org/10.1214/19-AOAS1265

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lymberopoulos, D., & McAuley, J. (2017). Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. *26th International Conference on*

World Wide Web. https://www.microsoft.com/en-us/research/publication/modeling-consumer-preferences-and-price-sensitivities-from-large-scale-grocery-shopping-transaction-logs/

A Appendix: Signal Functions

A.1 Small and Distinguishing Signal Functions

Here we flesh out the details of what it means for a signal function to be small and distinguishing. First, we introduce the grounded quadratic class of functions.

Definition 2. A function $g : \mathbb{N}_0^L \to \mathbb{R}$ is *grounded quadratic* if

$$g(\boldsymbol{q}) = \boldsymbol{b}'\boldsymbol{q} + \boldsymbol{q}'\boldsymbol{B}\boldsymbol{q}, \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{20}$$

for some $\boldsymbol{b} \in \mathbb{R}^L$ and some symmetric $L \times L$ matrix $\boldsymbol{B}$. The function g is *degenerate* if $g(\boldsymbol{q}) = 0$ for all $\boldsymbol{q} \in \mathbb{N}_0^L$.

The following condition helps ensure that the signal function does not pass through too much information about the true bundle.

Definition 3. A signal function S with support set $\mathcal{Q}$ is *distinguishing* if, for any non-degenerate grounded quadratic function g, there exists a set $Q \in \mathcal{Q}$ and $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q$ satisfying $g(\boldsymbol{q}) \neq g(\tilde{\boldsymbol{q}})$.

Our consistent theorem also requires signal functions satisfy the following further technical condition.

Definition 4. For a set Q let $|Q|$ denote the number of elements in Q. A signal function S with support $\mathcal{Q}$ is *small* if there exists a number $J \in \mathbb{N}$ so that

1. $|Q| \leq J$ for each $Q \in \mathcal{Q}$ and also
2. $||\boldsymbol{q}||_1 \leq J$ for all $\boldsymbol{q} \in \mathbb{N}_0^L$ such that $P\Big(S(\boldsymbol{q}) = \{\boldsymbol{q}\}\Big) < 1$.

A.2 The Small and Distinguishing Signal Function used in the Application

Assumption 3 requires the signal functions to be small and distinguishing. Here we provide an example of such a function which is the one we use in our application. In

what follows let J and J' be elements of $\mathbb{N}$. Let $\mathscr{L} = \{1, \ldots, L\}$. For each $\ell \in \mathscr{L}$ let I_ℓ be a subset of $\mathbb{N}$ which contains J' elements. For $\boldsymbol{q} \in \mathbb{N}_0^L$ let $Z(\boldsymbol{q}) \subseteq \mathscr{L}$ be defined by

$$Z(\boldsymbol{q}) = \{\ell \in \mathscr{L} : \boldsymbol{q}_\ell = 0\}$$

Similarly, define $N(\boldsymbol{q}) \subseteq \mathscr{L}$ by

$$N(\boldsymbol{q}) = \{\ell \in \mathscr{L} : \boldsymbol{q}_\ell \in I_\ell\}$$

We shall proceed to define an object $\tilde{S}(\boldsymbol{q})$ which will not contain $\boldsymbol{q}$ but will be defined so that $\{\boldsymbol{q}\} \cup \{\tilde{S}(\boldsymbol{q})\}$ is a signal function. To start, if $N(\boldsymbol{q}) = \varnothing$ or $\#Z(\boldsymbol{q}) < J$ then let $\tilde{S}(\boldsymbol{q}) = \varnothing$. On the other hand, if both $N(\boldsymbol{q}) \neq \varnothing$ and $\#Z(\boldsymbol{q}) \geq J$ then define $\tilde{S}(\boldsymbol{q})$ as follows.

Let $n(\boldsymbol{q})$ denote a uniform draw from $N(\boldsymbol{q})$. Let $z_1(\boldsymbol{q}), \ldots, z_J(\boldsymbol{q})$ denote uniform random draws from $Z(\boldsymbol{q})$ made without replacement. Further, let $i_j(\boldsymbol{q})$ denote a uniform draw from $I_{z_j(\boldsymbol{q})}$. Let $\pi^j(\boldsymbol{q}) = [\pi_1^j(\boldsymbol{q}), \ldots, \pi_L^j(\boldsymbol{q}))]$ denote the consumption bundle which satisfies

$$\pi_\ell^j(\boldsymbol{q}) = \begin{cases} i_j(\boldsymbol{q}), & \text{for } \ell = z_j(\boldsymbol{q}) \\ 0, & \text{for } \ell = n(\boldsymbol{q}) \\ q_\ell, & \text{for } \ell \notin \{z_j(\boldsymbol{q}), n(\boldsymbol{q})\} \end{cases}$$

Let $\tilde{S}(\boldsymbol{q})$ be uniform draws without replacement from $\{\pi^1(\boldsymbol{q}), \ldots, \pi^J(\boldsymbol{q})\}$ and let S be defined by

$$S(\boldsymbol{q}) = \{\boldsymbol{q}\} \cup \tilde{S}(\boldsymbol{q}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{21}$$

Proposition 7. *S defined in (21) is a small and distinguishing signal function.*

The proof of Proposition 7 is in Appendix B.

B Appendix: Proofs

B.1 Proof of Proposition 1

The proof of Proposition 1 relies on several lemmas. The first two lemmas are well-known and so we state them without proof. They concern the Gumbel and logistic distributions.

For $b \in \mathbb{R}$ write $X \sim \text{Gumbel}(b)$ if X is a random variable with CDF

$$F_X(x) = \exp\big(-\exp(b - x)\big)$$

The following lemma shows that the max of two independently distributed Gumbel random variables is itself a Gumbel random variable.

Lemma 1. *Suppose X and Y are independent and $X \sim$ Gumbel(b_X) and $Y \sim$ Gumbel(b_Y). Then*

$$\max(X, Y) \sim \textit{Gumbel}\Big(\ln(\exp(b_X) + \exp(b_Y))\Big)$$

Next, for $b \in \mathbb{R}$ write $X \sim \text{Logistic}(b)$ if X is a random variable with CDF

$$F_X(x) = \frac{1}{1 + \exp(b - x)}$$

The next lemma claims that the difference of two independent Gumbel random variables is a logistic random variable.

Lemma 2. *Let X and Y be independent and $X \sim$ Gumbel(b_X) and $Y \sim$ Gumbel(b_Y). Then*

$$Y - X \sim \textit{Logistic}(b_Y - b_X)$$

The next lemma establishes that a certain series is finite.

Lemma 3. *Suppose U is a standard Qua utility function. Then, for all $\boldsymbol{p} \in \mathbb{R}^L_{++}$,*

$$\sum_{\boldsymbol{q} \in \boldsymbol{N}_0^L} \exp\Big(U(\boldsymbol{q}, \boldsymbol{p})\Big) < \infty \tag{22}$$

The proof of Lemma 3 appears in subsection B.1.1. We now prove Proposition 1.

Proof of Proposition 1. Fix some consumption bundle $\boldsymbol{q} \in \mathbb{N}_0^L$ and some price vector $\boldsymbol{p} \in \mathbb{R}^L_{++}$. Let $\{\tilde{\boldsymbol{q}}^i\}_{i \in \mathbb{N}}$ be a non-repeating enumeration of the elements in $\mathbb{N}_0^L \backslash \{\boldsymbol{q}\}$. For $I \in \mathbb{N} \cup \{\infty\}$ let v_I be defined by

$$v_I = \ln\left(\sum_{i=1}^{I} \exp\Big(U(\tilde{\boldsymbol{q}}^i, \boldsymbol{p})\Big)\right) \tag{23}$$

and (recall that $\tilde{U}$ is defined in (5)) let $\tilde{v}_I$ be defined by

$$\tilde{v}_I = \sup_{i \leq I} \tilde{U}(\tilde{\boldsymbol{q}}^i, \boldsymbol{p}) \tag{24}$$

Repeated application of Lemma 1 gives

$$\tilde{v}_I \sim \text{Gumbel}\left(v_I\right), \qquad \text{for } I \in \mathbb{N} \tag{25}$$

By Theorem 13.4(i) in Billingsley (1986) $\tilde{v}_\infty$ is a random variable taking values in the extended real line. Let $a \in \mathbb{R}$. For $I \in \mathbb{N} \cup \{\infty\}$ let E_I be the event that $\tilde{v}_I$ is less than

or equal to a. We see that $E_I \downarrow E_\infty$ and using the continuity property of probability measures (see Billingsley (1986) Theorem 2.1) we see

$$P(E_\infty) = \lim_{I\to\infty} P(E_I) = \lim_{I\to\infty} \exp(-\exp(v_I - a)) = \exp(-\exp(v_\infty - a))$$

From Lemma 3 we know that $v_\infty < \infty$ and thus we see that

$$\tilde{v}_\infty \sim \text{Gumbel}(v_\infty)$$

Applying Lemma 2 we see

$$\begin{aligned}
f(\boldsymbol{q};\boldsymbol{p}) &= P\left(\tilde{U}(\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q}) \geq \max_{\tilde{\boldsymbol{q}}\in\mathbb{N}_0^L} \tilde{U}(\tilde{\boldsymbol{q}},\boldsymbol{p}'\tilde{\boldsymbol{q}})\right) \\
&= P\left(\tilde{U}(\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q}) \geq \tilde{v}_\infty\right) \\
&= P\left(0 \geq \tilde{v}_\infty - \tilde{U}(\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q})\right) \\
&= \frac{1}{1+\exp\left(v_\infty - U(\boldsymbol{A}\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q})\right)} \\
&= \frac{\exp(U(\boldsymbol{A}\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q}))}{\exp(U(\boldsymbol{A}\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q})) + \exp\left(v_\infty\right)} \\
&= \frac{\exp(U(\boldsymbol{A}\boldsymbol{q},\boldsymbol{p}'\boldsymbol{q}))}{\sum_{\tilde{\boldsymbol{q}}\in\mathbb{N}_0^L}\exp(U(\boldsymbol{A}\tilde{\boldsymbol{q}},\boldsymbol{p}'\tilde{\boldsymbol{q}}))}
\end{aligned}$$

which completes the proof. □

B.1.1 Proof of Lemma 3

The proof of Lemma 3 rests on a lemma which provides an upper bound on the number of elements in a certain set. So, for some set B let $|B|$ denote the number of elements in B. For some vector $\boldsymbol{v} = [v_1, \ldots, v_L] \in \mathbb{R}^L$ let $||\boldsymbol{v}||_1 = \sum_{\ell=1}^L |v_\ell|$. For $Q \in \mathbb{N}_0$, define the set $B(Q)$ by

$$B(Q) = \left\{\boldsymbol{q} \in \mathbb{N}_0^L : ||\boldsymbol{q}||_1 = Q\right\} \tag{26}$$

The following lemma gives a upper bound on $|B(Q)|$.

Lemma 4. *For $Q \in \mathbb{N}_0^L$,*

$$|B(Q)| \leq (Q+1)^{L-1} \tag{27}$$

Proof of Lemma 4. The number $|B(Q)|$ is the number of vectors $\boldsymbol{q} \in \mathbb{N}_0^L$ whose elements sum up to Q. This coincides with the number of ways in which Q identical balls can be

placed into L distinct boxes. We may thus find an expression for $|B(Q)|$ using the well-known formula for counting the number of ways of placing identical balls into distinct boxes and thus,

$$|B(Q)| = \binom{Q+L-1}{L-1}$$

Applying this formula we have

$$|B(Q)| = \frac{(Q+L-1)!}{(L-1)!Q!} = \prod_{\ell}^{L-1}\left(\frac{Q+\ell}{\ell}\right) \leq \prod_{\ell}^{L-1}(Q+1) = (Q+1)^{L-1}$$

which establishes (27). □

We now prove Lemma 3.

Proof of Lemma 3. Although the standard Qua utility function $U(\boldsymbol{q}, \boldsymbol{p})$ is only defined for positive price vectors $\boldsymbol{p} \gg 0$ we herein extend it to cover all price vectors $\boldsymbol{p} \in \mathbb{R}^L_+$ using the equation (4) as a definition. Now, given the quadratic form of U, the fact that U is strictly concave in $\boldsymbol{Aq}$ and U is strictly decreasing in expenditure, it is clear that there is some $\boldsymbol{q}^* \in \mathbb{N}^L$ and $u^* \in \mathbb{R}$ so that

$$u^* = U(\boldsymbol{q}^*, 0) \geq U(\boldsymbol{q}, \boldsymbol{p}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}^L_0 \text{ and } \boldsymbol{p} \in \mathbb{R}^L_+ \tag{28}$$

Fix some $\boldsymbol{p} \in \mathbb{R}^L_{++}$ and let $\rho > 0$ be the smallest element in $\boldsymbol{p}$. Clearly,

$$\boldsymbol{p}'\boldsymbol{q} \leq \rho||\boldsymbol{q}||_1 \tag{29}$$

Let $\boldsymbol{e}_1 = [1, 0, \ldots, 0]$. Now, applying (4), (28), the fact that $d_3 \geq 0$, and (29), we see

$$\begin{aligned} U(\boldsymbol{q}, \boldsymbol{p}) &= \boldsymbol{b}'\boldsymbol{Aq} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{Aq} - d_1\boldsymbol{p}'\boldsymbol{q} - d_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2d_3\boldsymbol{e}_1'\boldsymbol{Aq}\boldsymbol{p}'\boldsymbol{q} \\ &\leq u^* - d_1\boldsymbol{p}'\boldsymbol{q} - d_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2d_3\boldsymbol{e}_1'\boldsymbol{Aq}\boldsymbol{p}'\boldsymbol{q} \\ &\leq u^* - d_1\boldsymbol{p}'\boldsymbol{q} - d_2(\boldsymbol{p}'\boldsymbol{q})^2 \\ &\leq u^* - d_1\rho||\boldsymbol{q}||_1 - d_2\rho^2||\boldsymbol{q}||_1^2 \end{aligned} \tag{30}$$

Define $B(Q)$ by (26). Applying (30) and Lemma 4 we see

$$\begin{aligned}
\sum_{\boldsymbol{q}\in\boldsymbol{N}_0^L} \exp\Big(U(\boldsymbol{q},\boldsymbol{p})\Big) &\le \sum_{\boldsymbol{q}\in\boldsymbol{N}_0^L} \exp\Big(u^* - d_1\rho||\boldsymbol{q}||_1 - d_2\rho^2||\boldsymbol{q}||_1^2\Big) \\
&= \exp(u^*) \sum_{\boldsymbol{q}\in\boldsymbol{N}_0^L} \exp\Big(-d_1\rho||\boldsymbol{q}||_1 - d_2\rho^2||\boldsymbol{q}||_1^2\Big) \\
&= \exp(u^*) \sum_{Q=0}^{\infty} \sum_{\boldsymbol{q}\in B(Q)} \exp\Big(-d_1\rho||\boldsymbol{q}||_1 - d_2\rho^2||\boldsymbol{q}||_1^2\Big) \\
&= \exp(u^*) \sum_{Q=0}^{\infty} \sum_{\boldsymbol{q}\in B(Q)} \exp\Big(-d_1\rho Q - d_2\rho^2 Q^2\Big) \\
&\le \exp(u^*) \sum_{Q=0}^{\infty} (Q+1)^{L-1} \exp\Big(-d_1\rho Q - d_2\rho^2 Q^2\Big) < \infty
\end{aligned}$$

where the final inequality follows ffrom using the "ratio test" for the absolute convergence of series. □

B.2 Proof of Proposition 2

Let $\tilde{U}$ be a general Qua utility function with parameters $\boldsymbol{b}, \boldsymbol{A}, \boldsymbol{B}, d_1, d_2, \tilde{\boldsymbol{d}}$. As $\boldsymbol{B}$ is positive definite we know that both $\boldsymbol{B}^{1/2}$ and $\boldsymbol{B}^{-1/2}$ exist. Let $\boldsymbol{v}_1 \in \mathbb{R}^K$ be defined by

$$\boldsymbol{v}_1 = \frac{\boldsymbol{B}^{-1/2}\tilde{\boldsymbol{d}}}{||\boldsymbol{B}^{-1/2}\tilde{\boldsymbol{d}}||}$$

Next, let $\boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_K$ be K-vectors so that the $K \times K$ matrix $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_K]$ is orthogonal (i.e. $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$ where $\boldsymbol{I}$ is the $K \times K$ identity matrix).

Define $d_3 = \left\|\boldsymbol{B}^{-1/2}\tilde{\boldsymbol{d}}\right\| \ge 0$. From the definition of $\boldsymbol{V}$ we have

$$\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\boldsymbol{a} = d_3 a_1, \qquad \text{for all } \boldsymbol{a} = [a_1, a_2, \ldots, a_K] \in \mathbb{R}^K \tag{31}$$

Define an attribute matrix $\tilde{\boldsymbol{A}} = \boldsymbol{V}'\boldsymbol{B}^{1/2}\boldsymbol{A}$ and define $\tilde{U} : \mathbb{N}_0^L \times \mathbb{R}_{++}^L \to \mathbb{R}$ by

$$\tilde{U}(\boldsymbol{q},\boldsymbol{p}) = \boldsymbol{b}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q} - \boldsymbol{q}'\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\boldsymbol{q} - d_1(\boldsymbol{p}'\boldsymbol{q}) - d_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2d_3 a_1(\boldsymbol{p}'\boldsymbol{q}) \tag{32}$$

Clearly, this $\tilde{U}$ is a standard Qua utility function (note that $\boldsymbol{b}$ in (4) is equal to $\boldsymbol{V}'\boldsymbol{B}^{-1/2}\boldsymbol{b}$ in (32)).

We claim that the displayed equation in Proposition 2 holds. Let $\boldsymbol{e}_1 = [1, 0, 0, \ldots, 0] \in$

$\mathbb{R}^K$, $\boldsymbol{q} \in \mathbb{N}_0^L$, $\boldsymbol{p} \in \mathbb{R}_{++}^L$, $m = \boldsymbol{p}'\boldsymbol{q}$, and using (31) we see

$$\begin{aligned}
\tilde{U}(\boldsymbol{q},\boldsymbol{p}) &= \boldsymbol{b}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q} - \boldsymbol{q}'\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}^{1/2}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\tilde{\boldsymbol{A}}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{B}^{-1/2}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{q}m \\
&= \boldsymbol{b}'\boldsymbol{A}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{q} - d_1 m - d_2 m^2 - 2\tilde{\boldsymbol{d}}'\boldsymbol{A}\boldsymbol{q}m \\
&= U(\boldsymbol{q},\boldsymbol{p})
\end{aligned}$$

which proves the proposition.

B.3 Proof of Propositions 4 and 5

Proposition 4 is an obvious corollary of Proposition 5 and so only Proposition 5 is proved.

Fix some price vector $\boldsymbol{p}$. In what follows we use the shorthand $U(\boldsymbol{q}) \equiv U(\boldsymbol{q},\boldsymbol{p})$, $\boldsymbol{c} \equiv \boldsymbol{c}(\boldsymbol{p})$, and $f(\boldsymbol{q}) \equiv f(\boldsymbol{q}|\boldsymbol{p})$. We have

$$\begin{aligned}
\partial_{\boldsymbol{p}}\boldsymbol{E}[\boldsymbol{c}] &= \partial_{\boldsymbol{p}}\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L} f(\boldsymbol{q})\boldsymbol{q}\right) = \partial_{\boldsymbol{p}}\left(\frac{\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\boldsymbol{q}}{\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))}\right) \\
&= \frac{\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\right)\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\boldsymbol{q}\partial_{\boldsymbol{p}}U(\boldsymbol{q})\right)}{\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\right)^2} \\
&\quad - \frac{\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\boldsymbol{q}\right)\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\partial_{\boldsymbol{p}}U(\boldsymbol{q})\right)}{\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L}\exp(U(\boldsymbol{q}))\right)^2} \\
&= \left(\sum_{\boldsymbol{q}\in\mathbb{N}^L} f(\boldsymbol{q})\boldsymbol{q}\partial_{\boldsymbol{p}}U(\boldsymbol{q})\right) - \left(\sum_{\boldsymbol{q}\in\mathbb{N}^L} f(\boldsymbol{q})\boldsymbol{q}\right)\left(\sum_{\boldsymbol{q}\in\mathbb{N}^L} f(\boldsymbol{q})\partial_{\boldsymbol{p}}U(\boldsymbol{q})\right) \\
&= \boldsymbol{E}[\boldsymbol{c}\partial_{\boldsymbol{p}}U(\boldsymbol{c})] - (\boldsymbol{E}[\boldsymbol{c}])(\boldsymbol{E}[\partial_{\boldsymbol{p}}U(\boldsymbol{c})]) = \mathbf{Cov}\left(\boldsymbol{c}, \partial_{\boldsymbol{p}}U(\boldsymbol{c})\right)
\end{aligned}$$

which completes the proof.

B.4 Identification and Estimation Proofs

The following class of utility functions will prove useful.

Definition 5. A utility function $V : \mathbb{N}_0^L \times \mathbb{R}_{++}^L \to \mathbb{R}$ is a *quadratic expenditure modified utility function* (Quem utility function) if

$$V(\boldsymbol{q},\boldsymbol{p}) = \boldsymbol{\gamma}_0'\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{G}\boldsymbol{q} - \gamma_1(\boldsymbol{p}'\boldsymbol{q}) - \gamma_2(\boldsymbol{p}'\boldsymbol{q})^2 - 2\boldsymbol{\gamma}_3'\boldsymbol{q}(\boldsymbol{p}'\boldsymbol{q}) \tag{33}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^L$, $\boldsymbol{G}$ is a $L \times L$ symmetric matrix, γ_1, γ_2 are real numbers, and $\boldsymbol{\gamma}_3 \in \mathbb{R}^L$. A Quem V is degenerate if $V(\boldsymbol{q}, \boldsymbol{p}) = 0$ for all $\boldsymbol{q} \in \mathbb{N}_0^L$ and all $\boldsymbol{p} \in \mathbb{R}_{++}^L$.

Let $\boldsymbol{\gamma} = [\boldsymbol{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3]$ be a vector of parameters of the Quem utility function and let Γ be all such vectors. For $\boldsymbol{\gamma} \in \Gamma$ let $V(\cdot, \cdot; \boldsymbol{\gamma}) : \mathbb{N}_0^L \times \mathbb{R}_{++}^L \to \mathbb{R}$ be the Quem utility function with parameters $\boldsymbol{\gamma}$. For $\boldsymbol{\gamma} \in \Gamma$, non-empty $Q \subseteq \mathbb{N}_0^L$, and $\boldsymbol{p} \in \mathbb{R}_{++}^L$ let $h(\cdot|Q, \boldsymbol{p}; \boldsymbol{\gamma}) : \mathbb{N}_0^L \to [0, 1]$ be defined by

$$h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma}) = \frac{\exp(V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma}))}{\sum_{\tilde{\boldsymbol{q}} \in Q} \exp(V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \boldsymbol{\gamma}))}, \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{34}$$

where $V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma})$ is the Quem utility function with parameters $\boldsymbol{\gamma}$. Also, define $h(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\gamma})$ by

$$h(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\gamma}) = h(\boldsymbol{q}|\mathbb{N}_0^L, \boldsymbol{p}; \boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{35}$$

where $h(\boldsymbol{q}|\mathbb{N}_0^L, \boldsymbol{p}; \boldsymbol{\gamma})$ is defined by (34). The following lemma connects the Quem utility function to the standard Qua utility function.

Lemma 5. *Let $\psi : \Theta \to \Gamma$ be defined by (8). Then,*

$$U(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\theta}) = V(\boldsymbol{q}, \boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \textit{for all } \boldsymbol{q} \in \mathbb{N}_0^L,\ \boldsymbol{p} \in \mathbb{R}_{++}^L,\ \boldsymbol{\theta} \in \Theta \tag{36}$$

Consequently, if $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta})$ satisfies (7) then

$$f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}) = h(\boldsymbol{q}|\boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \textit{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{37}$$

Further, if $f(\cdot; Q, \boldsymbol{p}; \boldsymbol{\theta})$ satisfies (16) for some $Q \subseteq \mathbb{N}_0^L$ then

$$f(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\theta}) = h(\boldsymbol{q}|Q, \boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \textit{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{38}$$

Proof. The proof is just a matter of plugging in definitions. □

The following is crucial for the proof of Proposition 3.

Lemma 6. *Let h be defined by (34). Let S be a distinguishing signal function with support $\mathcal{Q}$, let $E \subseteq \mathbb{R}_{++}^L$ be a non-empty open set, and let $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \Gamma$. Then, $\boldsymbol{\gamma} \neq \tilde{\boldsymbol{\gamma}}$ if and only if there exists a non-empty open set $E' \subseteq E$ so that*

$$h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma}) \neq h(\boldsymbol{q}|Q, \boldsymbol{p}, \tilde{\boldsymbol{\gamma}}), \qquad \textit{for some } Q \in \mathcal{Q},\ \boldsymbol{q} \in Q,\ \textit{and all } \boldsymbol{p} \in E' \tag{39}$$

The proof of Lemma 6 is in Section B.4.1.

Proof of Proposition 3. It is trivial to show that *3.* $\implies$ *2.* $\implies$ *1.* So, we focus on the other direction. We show that *1.* $\implies$ *3.* or more precisely we show that not *3.* implies not *1.* So, assume that item *3.* does not hold.

Let ψ be defined by (8). As *3.* does not hold we see that $\psi(\boldsymbol{\theta}) \neq \psi(\tilde{\boldsymbol{\theta}})$. Next, let S be the signal function which satisfies $S(\boldsymbol{q}) = \mathbb{N}_0$ for all $\boldsymbol{q}$. It is easily shown that S is distinguishing. Thus, by setting $\boldsymbol{\gamma} = \psi(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\gamma}} = \psi(\tilde{\boldsymbol{\theta}})$ we may apply Lemma 6 to see that there is some non-empty open set $E' \subseteq E$ so that (39) holds. From Lemma 5 equation (36) we see that item *1.* does not hold. □

B.4.1 Proof of Lemma 6

The proof of Lemma 6 relies on 3 lemmas.

Lemma 7. *Let V be a non-degenerate Quem utility function and let $E \subseteq \mathbb{R}^L_+$ be a non-empty open set. There exists a $\boldsymbol{p} \in E$ so that $V(\boldsymbol{q}, \boldsymbol{p})$ is a non-degenerate grounded quadratic function of $\boldsymbol{q}$.*

Proof. Let V be a Quem utility function. Suppose that $V(\boldsymbol{q}, \boldsymbol{p})$ is a degenerate grounded quadratic function of $\boldsymbol{q}$ for all $\boldsymbol{p} \in E$. We shall show that V must be degenerate. First, note that

$$V(\boldsymbol{q}, \boldsymbol{p}) = 0 \qquad \text{for all } \ \boldsymbol{q} \in \mathbb{N}_0^L \ \text{ and all } \ \boldsymbol{p} \in E \tag{40}$$

Let us differentiate $V(\boldsymbol{q}, \boldsymbol{p})$ twice with respect to $\boldsymbol{p}$. This gives

$$\partial_{\boldsymbol{p}} V(\boldsymbol{q}, \boldsymbol{p}) = -\gamma_1 \boldsymbol{q} - 2\gamma_2 \boldsymbol{p}' \boldsymbol{q} \boldsymbol{q} - 2\boldsymbol{\gamma}_3' \boldsymbol{q} \boldsymbol{q} \tag{41}$$

$$\partial^2_{\boldsymbol{p}} V(\boldsymbol{q}, \boldsymbol{p}) = -2\gamma_2 \boldsymbol{q} \boldsymbol{q}' \tag{42}$$

Equation (40) implies that the expressions in both (41) and (42) are 0. From equation (42) we see that $\gamma_2 = 0$ (fix $\boldsymbol{q}$ and vary $\boldsymbol{p}$ to see this). Now, from (41) and the fact that $\gamma_2 = 0$ we see

$$-\gamma_1 - 2\boldsymbol{\gamma}_3' \boldsymbol{q} = 0, \qquad \text{for all } \boldsymbol{q} \neq 0$$

But, this is easily shown to imply $\gamma_1 = 0$ and $\boldsymbol{\gamma}_3 = 0$. So, we may now express $V(\boldsymbol{q}, \boldsymbol{p})$ as

$$V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) = \boldsymbol{\gamma}_0' \boldsymbol{q} - \boldsymbol{q}' \boldsymbol{G} \boldsymbol{q}$$

It is easy to show that $\boldsymbol{\gamma}_0 = 0$ and $\boldsymbol{G} = 0$ from (40). Thus, V is indeed degenerate. □

Lemma 8. *Let S be a distinguishing signal function with support $\mathcal{Q}$. Let $E \subseteq \mathbb{R}^L_{++}$ be a non-empty open set. For each non-degenerate Quem utility function V there exists a $Q \in \mathcal{Q}$, $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q$, and a non-empty open set $E' \subseteq E$ satisfying*

$$V(\boldsymbol{q}, \boldsymbol{p}) \neq V(\tilde{\boldsymbol{q}}, \boldsymbol{p}), \qquad \text{for all } \boldsymbol{p} \in E' \tag{43}$$

Proof. Let V be a non-degenerate Quem utility function. From Lemma 7 there exists a $\bar{\boldsymbol{p}} \in E$ so that $V(\boldsymbol{q}, \bar{\boldsymbol{p}})$ is a non-degenerate grounded quadratic function of $\boldsymbol{q}$. As S is distinguishing there exists a set $Q \in \mathcal{Q}$ and $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q$ satisfying

$$V(\boldsymbol{q}, \bar{\boldsymbol{p}}) \neq V(\tilde{\boldsymbol{q}}, \bar{\boldsymbol{p}}) \tag{44}$$

But, as V is a continuous function we may find some neighborhood of $\bar{\boldsymbol{p}}$ so that (44) holds for any $\boldsymbol{p}$ in this set. This establishes (43). □

Lemma 9. *Let (b_n) and $(\tilde{b}_n)$ be two sequences in $\mathbb{R}$ where $\sum_{n=1}^{\infty} \exp(b_n) < \infty$ and $\sum_{n=1}^{\infty} \exp(\tilde{b}_n) < \infty$. There exists a number $k \in \mathbb{R}$ so that $b_N = b_{\tilde{N}} + k$ for all $N \in \mathbb{N}$ if and only if*

$$\frac{\exp(b_N)}{\sum_{n=1}^{\infty} \exp(b_n)} = \frac{\exp(\tilde{b}_N)}{\sum_{n=1}^{\infty} \exp(\tilde{b}_n)}, \qquad \text{for each } N \in \mathbb{N} \tag{45}$$

Proof. The "only if" part is obvious so let's prove the "if" part. Let k be

$$k = \ln \left(\frac{\sum_{n=1}^{\infty} \exp(b_n)}{\sum_{n=1}^{\infty} \exp(\tilde{b}_n)} \right)$$

Rearranging (45) we see

$$\exp(b_N - \tilde{b}_N) = \frac{\sum_{n=1}^{\infty} \exp(b_n)}{\sum_{n=1}^{\infty} \exp(\tilde{b}_n)}$$

Taking logs of both sides and rearranging yields the result. □

We may now prove Lemma 6.

Proof of Lemma 6. The "if" part is obvious so we focus on the "only if" part. Suppose $\boldsymbol{\gamma} \neq \tilde{\boldsymbol{\gamma}}$. Let $W : \mathbb{N}_0^L \times \mathbb{R} \to \mathbb{R}$ be defined by

$$W(\boldsymbol{q}, \boldsymbol{p}) = V(\boldsymbol{q}, \boldsymbol{p}; \tilde{\boldsymbol{\gamma}}) - V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma})$$

It is easily shown that W is a non-degenerate Quem utility function. Thus, by Lemma 8 there exists $Q \in \mathcal{Q}$, $\boldsymbol{q}$, $\tilde{\boldsymbol{q}} \in Q$, and a non-empty open set $E' \subseteq E$ so that

$$W(\boldsymbol{q}, \boldsymbol{p}) \neq W(\tilde{\boldsymbol{q}}, \boldsymbol{p}), \qquad \text{for all } \boldsymbol{p} \in E'$$

Plugging in the definition of W and rearranging we see

$$V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \tilde{\boldsymbol{\gamma}}) - V(\boldsymbol{q}, \boldsymbol{p}; \tilde{\boldsymbol{\gamma}}) \neq V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \boldsymbol{\gamma}) - V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma})$$

This implies that for each $\boldsymbol{p} \in E'$ there is no constant $k \in \mathbb{R}$ so that both $V(\boldsymbol{q}, \boldsymbol{p}; \tilde{\boldsymbol{\gamma}}) = V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma}) + k$ and $V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \tilde{\boldsymbol{\gamma}}) = V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \boldsymbol{\gamma}) + k$. Thus, Lemma 9 establishes (39). □

B.4.2 Proof of Proposition 6

Fix some $\boldsymbol{q} \in \mathbb{N}_0^L$ and let $Q \in \mathcal{Q}$. Consider two cases: Case 1: $\boldsymbol{q} \notin Q$ and Case 2: $\boldsymbol{q} \in Q$. Suppose Case 1 holds. As the event $\boldsymbol{c} \in Q$ clearly implies $\boldsymbol{c} \neq \boldsymbol{q}$ we see that the right hand side of (15) is 0. On the other hand, by property (i) in the definition of a signal function, we see that the event $S(\boldsymbol{c}) = Q$ implies the event $\boldsymbol{c} \in Q$. But, we have already noted that the event $\boldsymbol{c} \in Q$ implies $\boldsymbol{c} \neq \boldsymbol{q}$ and so the left hand side of (15) is also 0. Thus, (15) holds under Case 1.

Now, consider Case 2. We write $\mathcal{S}$ for $S(\boldsymbol{c})$. By equation (14) and the assumption that $S(\boldsymbol{c})$ and $\boldsymbol{\rho}$ are independent conditional on $\boldsymbol{c}$

$$P\Big(\mathcal{S} = Q \Big| \boldsymbol{c} = \boldsymbol{q}, \boldsymbol{\rho}\Big) = P\Big(\mathcal{S} = Q \Big| \boldsymbol{c} = \tilde{\boldsymbol{q}}, \boldsymbol{\rho}\Big) \tag{46}$$

Now,

$$\begin{aligned}
P\Big(\boldsymbol{c} = \boldsymbol{q} \Big| \mathcal{S} = Q, \boldsymbol{\rho}\Big) &= \frac{P(\boldsymbol{c} = \boldsymbol{q} \text{ and } \mathcal{S} = Q | \boldsymbol{\rho})}{P(\mathcal{S} = Q | \boldsymbol{\rho})} \\
&= \frac{P(\boldsymbol{c} = \boldsymbol{q} \text{ and } \mathcal{S} = Q | \boldsymbol{\rho})}{\sum_{\tilde{\boldsymbol{q}} \in Q} P(\boldsymbol{c} = \tilde{\boldsymbol{q}} \text{ and } \mathcal{S} = Q | \boldsymbol{\rho})}, && \text{by prop (i) of def 1} \\
&= \frac{P(\boldsymbol{c} = \boldsymbol{q} | \boldsymbol{\rho}) P(S = Q | \boldsymbol{c} = \boldsymbol{q}, \boldsymbol{\rho})}{\sum_{\tilde{\boldsymbol{q}} \in Q} P(\boldsymbol{c} = \tilde{\boldsymbol{q}} | \boldsymbol{\rho}) P(S = Q | \boldsymbol{c} = \tilde{\boldsymbol{q}}, \boldsymbol{\rho})} \\
&= \frac{P(\boldsymbol{c} = \boldsymbol{q} | \boldsymbol{\rho})}{\sum_{\tilde{\boldsymbol{q}} \in Q} P(\boldsymbol{c} = \tilde{\boldsymbol{q}} | \boldsymbol{\rho})}, && \text{by (46)} \\
&= P\Big(\boldsymbol{c} = \boldsymbol{q} \Big| \boldsymbol{c} \in Q, \boldsymbol{\rho}\Big)
\end{aligned}$$

which establishes (15) for Case 2. Finally, (16) is a easy consequence of (15).

B.4.3 Proof of Theorem 1

Equation (19) implicitly treats the range of ψ, which is a list of matrices, vectors, and numbers, as a metric space. Let us be a bit more concrete about the metric space we have in mind. It is easily seen that the range of ψ is a subset of Γ which is the collection of all

lists of the form $\boldsymbol{\gamma} = [\boldsymbol{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3]$ where each such list corresponds to the parameters of the Quem utility function defined in (33). Let d be the metric on Γ defined by

$$d\Big([\boldsymbol{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3],[\tilde{\boldsymbol{G}}, \tilde{\boldsymbol{\gamma}}_0, \tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\boldsymbol{\gamma}}_3]\Big) \\ = d(\boldsymbol{G}, \tilde{\boldsymbol{G}}) + ||\boldsymbol{\gamma}_0 - \tilde{\boldsymbol{\gamma}}_0||_1 + |\gamma_1 - \tilde{\gamma}_1| + |\gamma_2 - \tilde{\gamma}_2| + ||\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}||_1$$

where $||\cdot||_1$ is the usual 1-norm and $d(\boldsymbol{G}, \tilde{\boldsymbol{G}}) = \sum_{j=1}^{L}\sum_{\ell=1}^{L} |\boldsymbol{G}_{j,\ell} - \tilde{\boldsymbol{G}}_{j,\ell}|$. We shall also use the same metric for Θ (that is, the distance between two lists in Θ is the sum of the distances between each element in the two lists). Because Γ and Θ are metric spaces we can meaningfully discuss open, closed, and compact subsets of these sets.

We also treat Γ as a vector space where, for real numbers k, k' we have

$$k[\boldsymbol{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3]+k'[\tilde{\boldsymbol{G}}, \tilde{\boldsymbol{\gamma}}_0, \tilde{\gamma}_1, \tilde{\gamma}_2, \tilde{\boldsymbol{\gamma}}_3] \\ = [k\boldsymbol{G} + k'\tilde{\boldsymbol{G}}, k\boldsymbol{\gamma}_0 + k'\tilde{\boldsymbol{\gamma}}_0, k\gamma_1 + k'\tilde{\gamma}_1, k\gamma_2 + k'\gamma_2, k\boldsymbol{\gamma}_3 + k'\tilde{\boldsymbol{\gamma}}_3]$$

Using the same approach we may treat Θ as a vector space. To prove Theorem 1 we require several lemmas.

Lemma 10. *Let $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, d_1, d_2, d_3] \in \Theta$ and suppose that $\tilde{\boldsymbol{A}}$ has rank K. For $\delta > 0$ let B_δ denote a closed ball in Γ centered at $\psi(\tilde{\boldsymbol{\theta}})$. There is a $\bar{\delta} > 0$ such that $\psi^{-1}(B_\delta)$ is compact for all $\delta \in (0, \bar{\delta}]$.*

Lemma 11. *Suppose Assumptions 3 and 5 hold and let h be defined by (35). Then,*

$$\boldsymbol{E}\left[\left|\ln\left(h(\boldsymbol{c}_n|\mathcal{S}_{n,i}, \boldsymbol{\rho}_n; \boldsymbol{\gamma})\right)\right|\right] < \infty, \qquad \text{for all } \boldsymbol{\gamma} \in \Gamma, \text{ and all } n, i \tag{47}$$

Lemma 12. *The function $\ln h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma})$ defined by (35) is concave in $\boldsymbol{\gamma} \in \Gamma$.*

Lemma 13. *Suppose $F_N(\boldsymbol{\gamma})$ is a random variable for each $N \in \mathbb{N}$ and $\boldsymbol{\gamma} \in \Gamma$. Suppose that F_N satisfies the following.*

1. *$F_N : \Gamma \to \mathbb{R}$ is concave for each $N \in \mathbb{N}$.*
2. *There exists a function $F : \Gamma \to \mathbb{R}$ so that $F_N(\boldsymbol{\gamma}) \xrightarrow{a.s.} F(\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \Gamma$.*

Then, F is concave and for any compact set $C \subseteq \Gamma$

$$\sup_{\boldsymbol{\gamma} \in C} \Big|F_N(\boldsymbol{\gamma}) - F(\boldsymbol{\gamma})\Big| \xrightarrow{a.s.} 0 \tag{48}$$

Lemma 14. *Let $F : \Gamma \to \mathbb{R}$ be a continuous function. Suppose there is a unique $\bar{\boldsymbol{\gamma}}$ which maximizes F. That is, there exists a $\bar{\boldsymbol{\gamma}} \in \Gamma$ so that*

$$F(\bar{\boldsymbol{\gamma}}) > F(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \neq \bar{\boldsymbol{\gamma}} \tag{49}$$

For $\delta > 0$ let B_δ be a closed ball in Γ centered at $\bar{\boldsymbol{\gamma}}$ with radius $\delta > 0$. For every compact set $C \subseteq \Gamma$ and every $\delta > 0$ there exists an $\varepsilon > 0$ so that

$$F(\bar{\boldsymbol{\gamma}}) - F(\boldsymbol{\gamma}) > \varepsilon, \qquad \text{for all } \boldsymbol{\gamma} \in C \backslash B_\delta \tag{50}$$

We now prove Theorem 1.

Proof of Theorem 1. For convenience let $\boldsymbol{\gamma}^* = \psi(\boldsymbol{\theta}^*)$. Let B_δ denote a closed ball in Γ of radius δ centered at $\boldsymbol{\gamma}^*$. By Assumption 4 and Lemma 10 there is a $\bar{\delta} > 0$ so that $\psi^{-1}(B_\delta)$ is compact for all $\delta \in (0, \bar{\delta}]$.

Next, define $Q_N : \Gamma \to \mathbb{R}$ by

$$Q_N(\boldsymbol{\gamma}) = \frac{1}{NI} \sum_{n=1}^{N} \sum_{i=1}^{I} \ln \Big(h(\boldsymbol{c}_n | \mathcal{S}_{i,n}, \boldsymbol{\rho}_n; \boldsymbol{\gamma}) \Big)$$

From Lemma 5 equation (38) we see

$$Q_N(\psi(\boldsymbol{\theta})) = \mathcal{L}_N(\boldsymbol{\theta}), \qquad \text{for all } \boldsymbol{\theta} \in \Theta \tag{51}$$

where $\mathcal{L}_N$ is defined by (17). Define $Q : \Gamma \to \mathbb{R}$ by

$$Q(\boldsymbol{\gamma}) = \boldsymbol{E} \left[\ln \Big(h(\boldsymbol{c}_1 | \mathcal{S}_{1,1}, \boldsymbol{\rho}_1; \boldsymbol{\gamma}) \Big) \right] \tag{52}$$

From Assumptions 3 and 5 and Lemma 11 we know $Q(\boldsymbol{\gamma})$ is finite for all $\boldsymbol{\gamma} \in \Gamma$. Thus, by Assumption 1 and the strong law of large numbers we know

$$Q_N(\boldsymbol{\gamma}) \xrightarrow{a.s.} Q(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \in \Gamma \tag{53}$$

From Lemma 12 it is clear that Q_N is concave and so we may use (53) and Lemma 13 to see

$$\sup_{\boldsymbol{\gamma} \in B_{\bar{\delta}}} \Big| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \Big| \xrightarrow{a.s.} 0 \tag{54}$$

From now on we conditional our analysis on a point in the sample space on which

$$\sup_{\boldsymbol{\gamma} \in B_{\bar{\delta}}} \Big| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \Big| \to 0 \tag{55}$$

We shall show that $\psi(\hat{\boldsymbol{\theta}}_N) \to \boldsymbol{\gamma}^*$. This will prove the theorem as the set of all points in the sample space which satisfy (55) contain a probability 1 event (by equation (54)).

Let $\delta \in (0, \bar{\delta})$ and for convenience let $\hat{\boldsymbol{\gamma}}_N = \psi(\hat{\boldsymbol{\theta}}_N)$. We shall show that there exists an $\bar{N} \in \mathbb{N}$ so that

$$||\hat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*|| < \delta, \quad \text{for all } N \geq \bar{N} \tag{56}$$

Of course (56) implies $\psi(\hat{\boldsymbol{\theta}}_N) \equiv \hat{\boldsymbol{\gamma}}_N \to \boldsymbol{\gamma}^*$ and so the proof is complete if we can show (56).

From Assumption 2, and Lemma 6, and the conditional Kullback-Leibler information inequality we see

$$Q(\boldsymbol{\gamma}^*) > Q(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \neq \psi(\boldsymbol{\theta}^*)$$

Thus, we may apply Lemma 14 to see that there exists an $\varepsilon > 0$ so that

$$Q(\boldsymbol{\gamma}^*) - Q(\boldsymbol{\gamma}) > \varepsilon, \qquad \text{for all } \boldsymbol{\gamma} \in B_{\bar{\delta}} \backslash B_\delta \tag{57}$$

From (55) there exists an $\bar{N} \in \mathbb{N}$ so that

$$\sup_{\boldsymbol{\gamma} \in B_{\bar{\delta}}} \left| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \right| < \tfrac{\varepsilon}{2}, \qquad \text{for all } N \geq \bar{N} \tag{58}$$

We claim that

$$Q_N(\boldsymbol{\gamma}^*) > Q_N(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \in B_{\bar{\delta}} \backslash B_\delta \quad \text{and all } N \geq \bar{N} \tag{59}$$

To see (59) use equations (57) and (58) and note that for all $\boldsymbol{\gamma} \in B_{\bar{\delta}} \backslash B_\delta$ and all $N \geq \bar{N}$

$$\begin{aligned} Q_N(\boldsymbol{\gamma}^*) - Q_N(\boldsymbol{\gamma}) &= Q_N(\boldsymbol{\gamma}^*) - Q(\boldsymbol{\gamma}^*) + Q(\boldsymbol{\gamma}^*) - Q(\boldsymbol{\gamma}) + Q(\boldsymbol{\gamma}) - Q_N(\boldsymbol{\gamma}) \\ &\geq Q(\boldsymbol{\gamma}^*) - Q_N(\boldsymbol{\gamma}) - |Q_N(\boldsymbol{\gamma}^*) - Q(\boldsymbol{\gamma}^*)| - |Q(\boldsymbol{\gamma}) - Q_N(\boldsymbol{\gamma})| \\ &> \varepsilon - \tfrac{\varepsilon}{2} - \tfrac{\varepsilon}{2} = 0 \end{aligned}$$

Thus, (59) holds. We next claim that

$$Q_N(\boldsymbol{\gamma}^*) > Q_N(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \notin B_\delta \quad \text{and all } N \geq \bar{N} \tag{60}$$

So, let $\boldsymbol{\gamma} \notin B_\delta$. If $\boldsymbol{\gamma} \in B_{\bar{\delta}}$ then (60) follows from (59) so suppose that $\boldsymbol{\gamma} \notin B_{\bar{\delta}}$. Let $t \in (0,1)$ be chosen so that $t\boldsymbol{\gamma} + (1-t)\boldsymbol{\gamma}^*$ is on the boundary of $B_{\bar{\delta}}$. Now, we use (59) and the concavity of Q_N to see

$$Q_N(\boldsymbol{\gamma}^*) > Q_N(t\boldsymbol{\gamma} + (1-t)\boldsymbol{\gamma}^*) \geq tQ_N(\boldsymbol{\gamma}) + (1-t)Q_N(\boldsymbol{\gamma}^*)$$

which gives (60) after rearranging.

B_δ is compact and so

$$\underset{\boldsymbol{\theta}\in B_\delta}{\operatorname{argmax}}\ \mathcal{L}_N(\boldsymbol{\theta}) \tag{61}$$

is non-empty. Further, from (60) we see that the argmax in (18) and (61) coincide. Thus, $\hat{\boldsymbol{\theta}}_N$ is an element of the argmax in (61) (for all $N \geq \bar{N}$) and consequently $||\hat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*|| \leq \delta$ for all $N \geq \bar{N}$. Thus, (56) holds and so the theorem has been proved. □

B.5 Proofs of lemmas used to prove Theorem 1

Proof of Lemma 10. For $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3] \in \Theta$ let $\boldsymbol{\theta}^1 = \boldsymbol{A}$, $\boldsymbol{\theta}^2 = \boldsymbol{b}$, $\boldsymbol{\theta}^3 = d_1$, $\boldsymbol{\theta}^4 = d_2$, and $\boldsymbol{\theta}^5 = d_3$. That is, $\boldsymbol{\theta}^1$ denotes the first entry in $\boldsymbol{\theta}$, $\boldsymbol{\theta}^2$ denotes the second entry, and so forth. Similarly, for $\boldsymbol{\gamma} = [\boldsymbol{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3] \in \Gamma$ let $\boldsymbol{\gamma}^1 = \boldsymbol{G}$, $\boldsymbol{\gamma}^2 = \boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}^3 = \gamma_1$, $\boldsymbol{\gamma}^4 = \gamma_2$, and $\boldsymbol{\gamma}^5 = \boldsymbol{\gamma_3}$.

Now, as $\tilde{\boldsymbol{A}}$ has rank K it is clear that $\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}$ also has rank K. Thus, there exists a closed ball C around $\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}$ which only contains matrices of rank K. Let $\bar{\delta} > 0$ be a number small enough so that $\boldsymbol{\gamma} \in B_{\bar{\delta}}$ implies $\boldsymbol{\gamma}^0 \in C$. Let $\delta \in (0, \bar{\delta}]$. We shall show that $\psi^{-1}(B_\delta)$ is compact.

As B_δ is closed and ψ is continuous we know that $\psi^{-1}(B_\delta)$ is closed. So, the desired conclusion holds if we can show that $\psi^{-1}(B_\delta)$ is bounded. Now, if $\psi^{-1}(B_\delta)$ were unbounded then there would be a sequence in $\psi^{-1}(B_\delta)$ denoted $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ where one of the vectors or matrices in $\boldsymbol{\theta}_n = [\boldsymbol{A}_n, \boldsymbol{b}_n, d_{1,n}, d_{2,n}, d_{3,n}]$ has an element whose absolute value goes to ∞. We shall show that this cannot be the case for any of the vectors nor the matrix in $[\boldsymbol{A}_n, \boldsymbol{b}_n, d_{1,n}, d_{2,n}, d_{3,n}]$.

First, it is clear that the sequence of matrices $\boldsymbol{A}_n$ could not have an element whose absolute value tends to infinity as we know $\boldsymbol{A}_n'\boldsymbol{A}_n \in C$. Second, as $\boldsymbol{A}_n$ is bounded we may assume, without loss of generality, that $\boldsymbol{A}_n$ converges to some matrix $\boldsymbol{A}$ (for if not just consider sub-sequences). From the definition of C, the fact that $\boldsymbol{A}_n'\boldsymbol{A}_n \in C$, and C is closed we see that each matrix $\boldsymbol{A}_n$ is rank K and additionally, $\boldsymbol{A}$ is rank K. Let $\lambda_n > 0$ be the smallest eigenvalue of $\boldsymbol{A}_n\boldsymbol{A}_n'$. As $\boldsymbol{A}$ is rank K (and so $\boldsymbol{A}\boldsymbol{A}'$ is also rank K) we know λ_n does not tend to 0. Now, we have

$$||\boldsymbol{A}_n'\boldsymbol{b}_n|| = \sqrt{\boldsymbol{b}_n'\boldsymbol{A}_n\boldsymbol{A}_n'\boldsymbol{b}_n} \geq \lambda_n||\boldsymbol{b}_n||$$

Thus, if $\boldsymbol{b}_n$ has an entry which tends to infinity in absolute value then also $||\boldsymbol{A}_n'\boldsymbol{b}_n|| \to \infty$ which contradicts $\boldsymbol{\theta}_n \in \psi^{-1}(B_\delta)$.

Next, note that neither $|d_{1,n}|$ nor $|d_{2,n}|$ tend to infinity as a direct result of $\boldsymbol{\gamma}_n \in \psi^{-1}(B_\delta)$. Finally, we have

$$||d_{3,n}\boldsymbol{A}_n'\boldsymbol{e}_1|| = |d_{3,n}|\sqrt{\boldsymbol{e}_1'\boldsymbol{A}_n\boldsymbol{A}_n'\boldsymbol{e}_1} \geq |d_{3,n}|\lambda_n$$

So, if $|d_{3,n}| \to \infty$ then $||d_{3,n}\boldsymbol{A}_n'\boldsymbol{e}_1||$ also goes to infinity which contradicts $\boldsymbol{\theta}_n \in \psi^{-1}(B_\delta)$. □

Proof of Lemma 12. It is clear that $V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma})$, defined by (33), is linear in parameters and so it can be represented as

$$V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma}) = w(\boldsymbol{q}, \boldsymbol{p})'\boldsymbol{\gamma}$$

where $w(\boldsymbol{q}, \boldsymbol{p})$ is some transformation of the data. Thus, $\ln h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma})$ can be expressed as

$$\ln h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma}) = \ln\left(\frac{\exp(w(\boldsymbol{q}, \boldsymbol{p})'\boldsymbol{\gamma})}{\sum_{\tilde{\boldsymbol{q}}\in Q}\exp(w(\tilde{\boldsymbol{q}}, \boldsymbol{p})'\boldsymbol{\gamma})}\right)$$

But, this is the form of the log likelihood of the conditional logit and it is well-known that this function is concave in $\boldsymbol{\gamma}$. See for example McFadden (1974). □

Proof of Lemma 11. Let $\boldsymbol{\gamma} \in \Gamma$. Using Assumption 3 it is straightforward to show that there is some $M \in \mathbb{R}$ large enough so that for all $\boldsymbol{q} \in \mathbb{N}_0^L$, $\tilde{\boldsymbol{q}} \in S(\boldsymbol{q})$, and all $\boldsymbol{p} \in \mathbb{R}_{++}^L$

$$|V(\tilde{\boldsymbol{q}}, \boldsymbol{p}; \boldsymbol{\gamma})| \leq \left[\sum_{k=0}^{2}\sum_{j=0}^{2}||\boldsymbol{q}||_k^k||\boldsymbol{p}||_j^j\right] M$$

Again applying Assumption 3 we see

$$\begin{aligned}
0 \leq -\ln\left(h(\boldsymbol{c}_n|\mathcal{S}_{n,i}, \boldsymbol{\rho}_n; \boldsymbol{\gamma})\right) &= -\ln\left(\frac{\exp(V(\boldsymbol{c}_n, \boldsymbol{\rho}_n; \boldsymbol{\gamma}))}{\sum_{\tilde{\boldsymbol{c}}\in\mathcal{S}_{n,i}}\exp(V(\tilde{\boldsymbol{c}}, \boldsymbol{\rho}_n; \boldsymbol{\gamma}))}\right) \\
&\leq -\ln\left(\frac{\exp\left(-\left[\sum_{k=0}^{2}\sum_{j=0}^{2}||\boldsymbol{c}_n||_k^k||\boldsymbol{\rho}_n||_j^j\right] M\right)}{J\exp\left(\left[\sum_{k=0}^{2}\sum_{j=0}^{2}||\boldsymbol{c}_n||_k^k||\boldsymbol{\rho}_n||_j^j\right] M\right)}\right) \\
&= 2\left[\sum_{k=0}^{2}\sum_{j=0}^{2}||\boldsymbol{c}_n||_k^k||\boldsymbol{\rho}_n||_j^j\right] M + \ln J
\end{aligned}$$

Using this inequality and Assumption 5 we have

$$\boldsymbol{E}\Big[\big|\ln\left(h(\boldsymbol{c}_n|\mathcal{S}_{n,i}, \boldsymbol{\rho}_n; \boldsymbol{\gamma})\right)\big|\Big] \leq 2\left[\sum_{k=0}^{2}\sum_{j=0}^{2}\boldsymbol{E}\Big[||\boldsymbol{c}_n||_k^k||\boldsymbol{\rho}_n||_j^j\Big]\right] M + \ln J < \infty$$

which is what we needed to show. □

Proof of Lemma 13. Let $\tilde{\Gamma} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots\}$ be some countable dense subset of Γ. By item 2 of the lemma there is some probability 1 event E so that $F_N(\boldsymbol{\gamma}_k) \to F(\boldsymbol{\gamma}_k)$ for all $k \in \mathbb{N}$ as $N \to \infty$ on E. Let $C \subseteq \Gamma$ be compact. By Rockafellar (1970) Theorem 10.8 and item 1 of the lemma, the function F is concave and the function F_N converges uniformly on C to F for any sample space point in E. But, E is a probability 1 event and so (48) holds. □

Proof of Lemma 14. Suppose $C \subseteq \Gamma$ is compact and suppose, for a contradiction, that there is some $\delta > 0$ so that there is no $\varepsilon > 0$ so that (50) holds. Then, we can find a convergent sequence $\boldsymbol{\gamma}_n$ in C where $F(\boldsymbol{\gamma}_n) \to F(\bar{\boldsymbol{\gamma}})$ and each element of the sequence satisfies $||\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_n|| > \delta$. From the continuity of F the limit of the sequence would contradict the condition in (49). □

B.6 Proof of Proposition 7

Lemma 15. *If S defined by* (21) *is a signal function then it is distinguishing.*

The proof of Lemma 15 follows the proof of Proposition 7.

Proof of Proposition 7. It is clear that S satisfies (i) and (iii) of Definition 1. So, we show that S satisfies (ii). Now, if $\boldsymbol{q} \in \mathbb{N}_0^L$ has $N(\boldsymbol{q} = \varnothing)$ or $\#Z(\boldsymbol{q}) < J$ then $S(\boldsymbol{q}) = \{\boldsymbol{q}\}$ and so (14) holds. Now, suppose $\boldsymbol{q} \in \mathbb{N}_0^L$ is such that $N(\boldsymbol{q}) \neq \varnothing$ and $\#Z(\boldsymbol{q}) \geq J$. Let Q be some set in $\mathcal{Q}(\boldsymbol{q})$. Let $\tilde{\boldsymbol{q}} \in Q$. First note that $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q$ implies $\#Z(\boldsymbol{q}) = \#Z(\tilde{\boldsymbol{q}})$ and $\#N(\boldsymbol{q}) = \#N(\tilde{\boldsymbol{q}})$. Now, it is easy to show

$$\begin{aligned} P(S(\boldsymbol{q}) = Q) &= \frac{1}{\#N(\boldsymbol{q})\#Z(\boldsymbol{q})[\#Z(\boldsymbol{q}) - 1] \ldots [\#Z(\boldsymbol{q}) - J]} \\ &= \frac{1}{\#N(\tilde{\boldsymbol{q}})\#Z(\tilde{\boldsymbol{q}})[\#Z(\tilde{\boldsymbol{q}}) - 1] \ldots [\#Z(\tilde{\boldsymbol{q}}) - J]} = P(S(\tilde{\boldsymbol{q}}) = Q) \end{aligned}$$

Thus, S is a signal function. It is obvious that S is small so we just need to show that it is distinguishing. That is S is distinguishing follows from Lemma 15. □

B.7 Proof of Lemma 15

Let B be an $M \times M$ symmetric matrix and let $b_{i,j}$ denote the entry in row i, column j. Let $\text{diag}(B) = [b_{1,1}, \ldots, b_{M,M}] \in \mathbb{R}^M$. That is, $\text{diag}(B)$ is the vector composed of the elements on the diagonal of B. Similarly, let $\text{off}(B)$ denote the vector

$$\text{off}(B) = [b_{i,j}]_{i<j} = [b_{1,2}, b_{1,3}, \ldots, b_{M-1,M}]$$

In other words, $\text{off}(B) \in \mathbb{R}^{M(M-1)/2}$ denotes the off-diagonal elements of the bottom-left triangle of matrix B.

Let S be a signal function. Let $\mathcal{Q}_1(\boldsymbol{q}) = \{Q \subseteq \mathbb{N}_0^L : P(Q \in S(\boldsymbol{q})) > 0\}$. Let $\mathcal{Q}_2(\boldsymbol{q})$ be defined by

$$\mathcal{Q}_2(\boldsymbol{q}) = \bigcup_{\tilde{\boldsymbol{q}} \in \mathcal{Q}(\boldsymbol{q})} \mathcal{Q}_1(\tilde{\boldsymbol{q}})$$

In other words, $\mathcal{Q}_2(\boldsymbol{q})$ is the union of the supports of $S(\tilde{\boldsymbol{q}})$ where the union is taken over all $\tilde{\boldsymbol{q}} \in \mathcal{Q}(\boldsymbol{q})$. Note that because $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{q})$ we have $\mathcal{Q}_1(\boldsymbol{q}) \subseteq \mathcal{Q}_2(\boldsymbol{q})$. Similarly, for all $n \in \mathbb{N}$ define

$$\mathcal{Q}_{n+1}(\boldsymbol{q}) = \bigcup_{\tilde{\boldsymbol{q}} \in \mathcal{Q}(\boldsymbol{q})} \mathcal{Q}_n(\tilde{\boldsymbol{q}})$$

Finally, let $\mathcal{Q}_\infty(\boldsymbol{q}) = \cup_{n \in \mathbb{N}} \mathcal{Q}_n(\boldsymbol{q})$. It is easy to verify that $\tilde{\boldsymbol{q}} \in \mathcal{Q}_\infty(\boldsymbol{q})$ if and only if there exists a sequence $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N$ so that

$$\boldsymbol{q}_1 \in \mathcal{Q}(\boldsymbol{q}), \quad \boldsymbol{q}_2 \in \mathcal{Q}(\boldsymbol{q}_1), \quad \boldsymbol{q}_3 \in \mathcal{Q}(\boldsymbol{q}_2), \quad \ldots, \quad \boldsymbol{q}_N \in \mathcal{Q}(\boldsymbol{q}_{N-1}), \quad \text{and} \quad \tilde{\boldsymbol{q}} \in \mathcal{Q}(\boldsymbol{q}_N)$$

For a set $Q \subseteq \mathbb{N}_0^L$ let $1_Q(\boldsymbol{q})$ be defined by

$$1_Q(\boldsymbol{q}) = \begin{cases} 1, & \text{if } \boldsymbol{q} \in Q \\ 0, & \text{else} \end{cases}$$

Lemma 16. *Let S be a signal function and let $\tilde{Q} = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N\}$ be a finite subset of $\mathbb{N}_0^L$. Let $\tilde{\mathcal{Q}}$ be defined by*

$$\tilde{\mathcal{Q}} = \bigcup_{\boldsymbol{q} \in \tilde{Q}} \mathcal{Q}_\infty(\boldsymbol{q})$$

Let $\{Q_1, \ldots, Q_M\}$ be an arbitrary enumeration of the elements in $\tilde{\mathcal{Q}}$. If the matrix

$$D = \begin{bmatrix} \boldsymbol{q}_1' & \text{diag}(\boldsymbol{q}_1\boldsymbol{q}_1') & \text{off}(\boldsymbol{q}_1\boldsymbol{q}_1') & 1_{Q_1}(\boldsymbol{q}_1) & \ldots & 1_{Q_M}(\boldsymbol{q}_1) \\ & & \cdots & & & \\ \boldsymbol{q}_N' & \text{diag}(\boldsymbol{q}_N\boldsymbol{q}_N') & \text{off}(\boldsymbol{q}_N\boldsymbol{q}_N') & 1_{Q_1}(\boldsymbol{q}_N) & \ldots & 1_{Q_M}(\boldsymbol{q}_N) \end{bmatrix}$$

has full column rank then S is distinguishing.

Proof. We prove the lemma by supposing that S is not distinguishing and then we show that the matrix D does not have full column rank. So, suppose S is not distinguishing. This means that there is a $\boldsymbol{b} \in \mathbb{R}^L$ and a symmetric $L \times L$ matrix B where, for all $Q \in \tilde{\mathcal{Q}}$

$$\boldsymbol{b}'\boldsymbol{q} + \boldsymbol{q}'B\boldsymbol{q} = \boldsymbol{b}'\tilde{\boldsymbol{q}} + \tilde{\boldsymbol{q}}'B\tilde{\boldsymbol{q}}, \qquad \text{for all } \boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q \tag{62}$$

and it is not the case that both $\boldsymbol{b} = 0$ and $B = 0$. Note that (62) implies that for each $m \in \{1, \ldots, M\}$ there exists a number β_m so that

$$\beta_m = \boldsymbol{b}'\boldsymbol{q} + \boldsymbol{q}' B \boldsymbol{q}, \qquad \text{for all } \boldsymbol{q} \in Q_m$$

This is clearly implies

$$0 = \boldsymbol{b}'\boldsymbol{q}_n + \boldsymbol{q}_n' B \boldsymbol{q}_n - \sum_{m=1}^{M} \beta_m 1_{Q_m}(\boldsymbol{q}_n), \qquad \text{for all } n \in \{1, \ldots, N\} \tag{63}$$

Let $\boldsymbol{v}$ be the vector defined by

$$\boldsymbol{v} = [\boldsymbol{b}', \operatorname{diag}(B)', 2\operatorname{off}(B)', -\beta_1, \ldots, -\beta_M]$$

Applying (63) we see

$$\begin{aligned} D\boldsymbol{v} &= \begin{bmatrix} \boldsymbol{b}'\boldsymbol{q}_1 + \operatorname{diag}(B)' \operatorname{diag}(\boldsymbol{q}_1\boldsymbol{q}_1') + 2\operatorname{off}(B)' \operatorname{off}(\boldsymbol{q}_1\boldsymbol{q}_1') - \sum_{m=1}^{M} \beta_m 1_{Q_m}(\boldsymbol{q}_1) \\ \cdots \\ \boldsymbol{b}'\boldsymbol{q}_N + \operatorname{diag}(B)' \operatorname{diag}(\boldsymbol{q}_N\boldsymbol{q}_N') + 2\operatorname{off}(B)' \operatorname{off}(\boldsymbol{q}_N\boldsymbol{q}_N') - \sum_{m=1}^{M} 1_{Q_m}(\boldsymbol{q}_N) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{b}'\boldsymbol{q}_1 + \boldsymbol{q}_1' B \boldsymbol{q}_1 - \sum_{m=1}^{M} \beta_m 1_{Q_m}(\boldsymbol{q}_1) \\ \cdots \\ \boldsymbol{b}'\boldsymbol{q}_N + \boldsymbol{q}_N' B \boldsymbol{q}_N - \sum_{m=1}^{M} 1_{Q_m}(\boldsymbol{q}_N) \end{bmatrix} = \boldsymbol{0} \end{aligned}$$

Thus, D cannot have full column rank. □

We can now prove Lemma 15.

Proof of Lemma 15. Let $v_{1,\ell}, v_{2,\ell}, \ldots, v_{J,\ell}$ enumerate the values in I_ℓ. Let $\tilde{L} = L(L-1)/2$. Let $\tilde{Q}_1$ denote the finite subset of $\mathbb{N}_0^L$ which satisfies

$$\tilde{Q}_1 = \left\{ v_{1,\ell} \boldsymbol{e}_\ell \in \mathbb{N}_0^L : \ell \in \{1, \ldots, L\} \right\} \equiv \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L\}$$

Let $\tilde{Q}_2$ be defined by

$$\tilde{Q}_2 = \left\{ v_{2,\ell} \boldsymbol{e}_\ell \in \mathbb{N}_0^L : \ell \in \{1, \ldots, L\} \right\} \equiv \{\boldsymbol{q}_{L+1}, \ldots, \boldsymbol{q}_{2L}\}$$

Let $\tilde{Q}_{11}$ be defined by

$$\tilde{Q}_{11} = \left\{ v_{1,\ell} \boldsymbol{e}_\ell + v_{1,k} \boldsymbol{e}_k \in \mathbb{N}_0^L : \ell \neq k \right\} \equiv \{\boldsymbol{q}_{2L+1}, \ldots \boldsymbol{q}_{\tilde{L}}\}$$

Let $\tilde{Q}_3$ be the singleton set defined by

$$\tilde{Q}_3 = \{v_{3,1} \boldsymbol{e}_1\} \equiv \{\boldsymbol{q}_{\tilde{L}+1}\}$$

Let $\tilde{Q}_{12}$ be the singleton set defined by

$$\tilde{Q}_{12} = v_{1,1}\boldsymbol{e}_1 + v_{2,2}\boldsymbol{e}_2 \equiv \{\boldsymbol{q}_{\tilde{L}+2}\}$$

Let $\tilde{Q}$ be the union of the sets just defined. That is,

$$\tilde{Q} = \tilde{Q}_1 \cup \tilde{Q}_2 \cup \tilde{Q}_3 \cup \tilde{Q}_{11}$$

It is clear that for all $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in \tilde{Q}_1 \cup \tilde{Q}_2 \cup \tilde{Q}_3$ we have $\mathcal{Q}_\infty(\boldsymbol{q}) = \mathcal{Q}_\infty(\tilde{\boldsymbol{q}})$. Let Q_1 denote this subset of $\mathbb{N}_0^L$ (so, $Q_1 = \mathcal{Q}_\infty(\boldsymbol{q})$ for any $\boldsymbol{q} \in \tilde{Q}_1 \cup \tilde{Q}_2 \cup \tilde{Q}_3$). Similarly, it is clear that for all $\boldsymbol{q}, \tilde{\boldsymbol{q}} \in \tilde{Q}_{11} \cup \tilde{Q}_{12}$ we have $\mathcal{Q}_\infty(\boldsymbol{q}) = \mathcal{Q}_\infty(\tilde{\boldsymbol{q}})$. Let Q_2 denote this subset of $\mathbb{N}_0^L$ (so, $Q_2 = \mathcal{Q}_\infty(\boldsymbol{q})$ for any $\boldsymbol{q} \in \tilde{Q}_{11} \cup \tilde{Q}_{12}$).

We shall apply Lemma 16. To do this, let D denote the matrix in Lemma 16. Let V_1 denote the $L \times L$ diagonal matrix whose diagonal entries are $v_{1,1}, v_{1,2}, \ldots, v_{1,L}$. Let V_2 be the $L \times L$ diagonal matrix whose diagonal entries are $v_{2,1}, v_{2,2}, \ldots, v_{2,L}$. Let V_3 denote the $\tilde{L} \times \tilde{L}$ diagonal matrix whose diagonal entries are $v_{1,\ell}v_{1,k}$ for $\ell < k$ where the entries may be enumerated as $v_{1,1}v_{1,2},\ \ldots,\ v_{1,1}v_{1,L},\ v_{1,2}v_{1,3},\ \ldots v_{1,2}v_{1,L},\ v_{1,3}v_{1,4},\ \ldots, v_{1,L-1}v_{1,L}$. If D and D' are two matrices with the same rank then write $D \sim D'$. We will alter D using elementary row and column operations (which preserve the rank). It can be verified that D can be expressed as

$$D = \begin{bmatrix} V_1 & V_1^2 & 0 & \iota & 0 \\ V_2 & V_2^2 & 0 & \iota & 0 \\ X_1 & X_2 & V_3 & 0 & \iota \\ v_{3,1}\boldsymbol{e}_1' & v_{3,1}^2\boldsymbol{e}_1' & 0 & 1 & 0 \\ \boldsymbol{x}_1' & \boldsymbol{x}_2' & v_{1,1}v_{2,2}\boldsymbol{e}_1' & 0 & 1 \end{bmatrix}$$

First, we perform a column and row swap which yields

$$D \sim \begin{bmatrix} V_1 & V_1^2 & \iota & 0 & 0 \\ V_2 & V_2^2 & \iota & 0 & 0 \\ v_{3,1}\boldsymbol{e}_1' & v_{3,1}^2\boldsymbol{e}_1' & 1 & 0 & 0 \\ X_1 & X_2 & 0 & V_3 & \iota \\ \boldsymbol{x}_1' & \boldsymbol{x}_2' & 0 & v_{1,1}v_{2,2}\boldsymbol{e}_1' & 1 \end{bmatrix} \equiv D'$$

Note that the matrix D' has the same rank as D and because of the presence of the zeros in the upper right of D' we see that D' has full rank if the following two matrices are each full rank

$$D_1 = \begin{bmatrix} V_1 & V_1^2 & \iota \\ V_2 & V_2^2 & \iota \\ v_{3,1}\boldsymbol{e}_1' & v_{3,1}^2\boldsymbol{e}_1' & 1 \end{bmatrix} \qquad \text{and} \qquad D_2 = \begin{bmatrix} V_3 & \iota \\ v_{1,1}v_{2,2}\boldsymbol{e}_1' & 1 \end{bmatrix}$$

Now, we have

$$
\begin{aligned}
D_1 &\sim \begin{bmatrix} I & V_1 & V_1^{-1}\iota \\ V_2 & V_2^2 & \iota \\ v_{3,1}\boldsymbol{e}_1' & v_{3,1}^2\boldsymbol{e}_1' & 1 \end{bmatrix} \sim \begin{bmatrix} I & V_1 & V_1^{-1}\iota \\ 0 & V_2^2 - V_2V_1 & \iota - V_2V_1^{-1}\iota \\ 0 & (v_{3,1}^2 - v_{3,1}v_{1,1})\boldsymbol{e}_1' & 1 - \frac{v_{3,1}}{v_{1,1}} \end{bmatrix} \\
&\sim \begin{bmatrix} I & V_1 & V_1^{-1}\iota \\ 0 & I & (V_2 - V_1)^{-1}(V_2^{-1} - V_1^{-1})\iota \\ 0 & (v_{3,1}^2 - v_{3,1}v_{1,1})\boldsymbol{e}_1' & 1 - \frac{v_{3,1}}{v_{1,1}} \end{bmatrix} \\
&= \begin{bmatrix} I & V_1 & V_1^{-1}\iota \\ 0 & I & -V_2^{-1}V_1^{-1}\iota \\ 0 & (v_{3,1}^2 - v_{3,1}v_{1,1})\boldsymbol{e}_1' & 1 - \frac{v_{3,1}}{v_{1,1}} \end{bmatrix} \\
&\sim \begin{bmatrix} I & V_1 & V_1^{-1}\iota \\ 0 & I & V_2^{-1}V_1^{-1}\iota \\ 0 & 0 & 1 - \frac{v_{3,1}}{v_{1,1}} + \frac{v_{3,1}^2 - v_{3,1}v_{1,1}}{v_{1,1}v_{2,1}} \end{bmatrix}
\end{aligned}
$$

So, D_1 is full rank if $1 - \frac{v_{3,1}}{v_{1,1}} + \frac{v_{3,1}^2 - v_{3,1}v_{1,1}}{v_{1,1}v_{2,1}} \neq 0$. This is equivalent to

$$v_{1,1}v_{2,1} - v_{2,1}v_{3,1} + v_{3,1}^2 - v_{1,1}v_{3,1} = 0$$

We have

$$v_{1,1}v_{2,1} - v_{2,1}v_{3,1} + v_{3,1}^2 - v_{1,1}v_{3,1} = (v_{3,1} - v_{1,1})(v_{3,1} - v_{1,2})$$

and so D_1 is full rank unless $v_{3,1} = v_{1,1}$ or $v_{3,1} = v_{2,1}$. But, from the definition of $v_{1,1}$, $v_{2,1}$, and $v_{3,1}$ we know this is not the case. Thus, D_1 is full rank. We now show that D_2 is full rank.

$$D_2 \sim \begin{bmatrix} I & V_3^{-1}\iota \\ v_{1,1}v_{2,2}\boldsymbol{e}_1' & 1 \end{bmatrix} \sim \begin{bmatrix} I & V_3^{-1}\iota \\ 0 & 1 - \frac{v_{1,1}v_{2,2}}{v_{1,1}v_{1,2}} \end{bmatrix} = \begin{bmatrix} I & V_3^{-1}\iota \\ 0 & 1 - \frac{v_{2,2}}{v_{1,2}} \end{bmatrix} \tag{64}$$

Thus, D_2 is full rank if $v_{1,2} \neq v_{2,2}$. But, this is true from the definition of $v_{1,2}$ and $v_{2,2}$ and so D_2 is full rank. So, we see that D is full rank and so, by Lemma 16, S is distinguishing. □

B.8 Choice of the dimensionality, K

The fitted $\hat{\boldsymbol{A}}$ can be viewed as L observations of a multivariate random variable, whose empirical sample covariance is $\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'$. Applying principal component analysis to $\hat{\boldsymbol{A}}$ rotates the embedding coordinates such that its empirical covariance is diagonal. In our empirical application, the contribution of the smallest variance in this diagonalization would appear to be fairly low, measuring only 1.8% of the sum of the variances. This is illustrated in Figure 12.

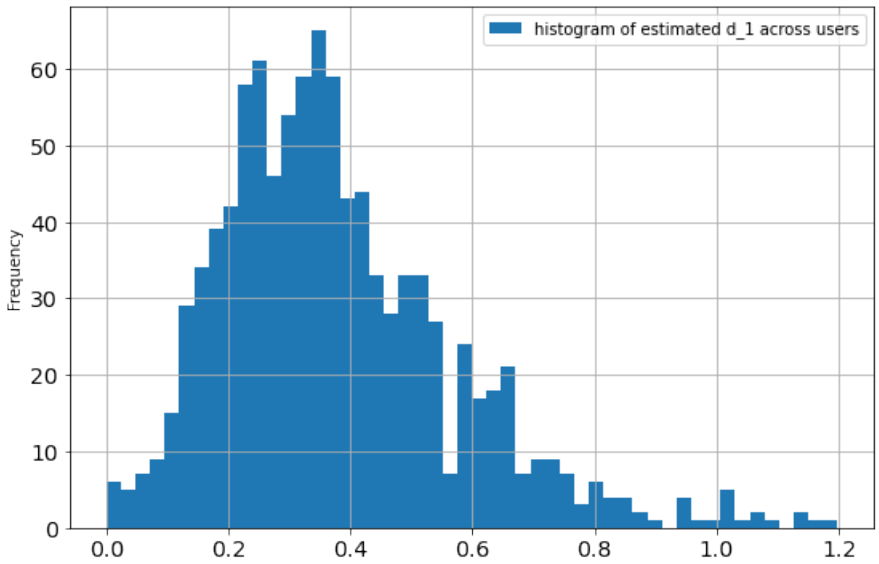

Figure 7: A histogram of the estimated values of d_1 for the 986 tracked customers in the DunnHumby dataset as described in Section 5.

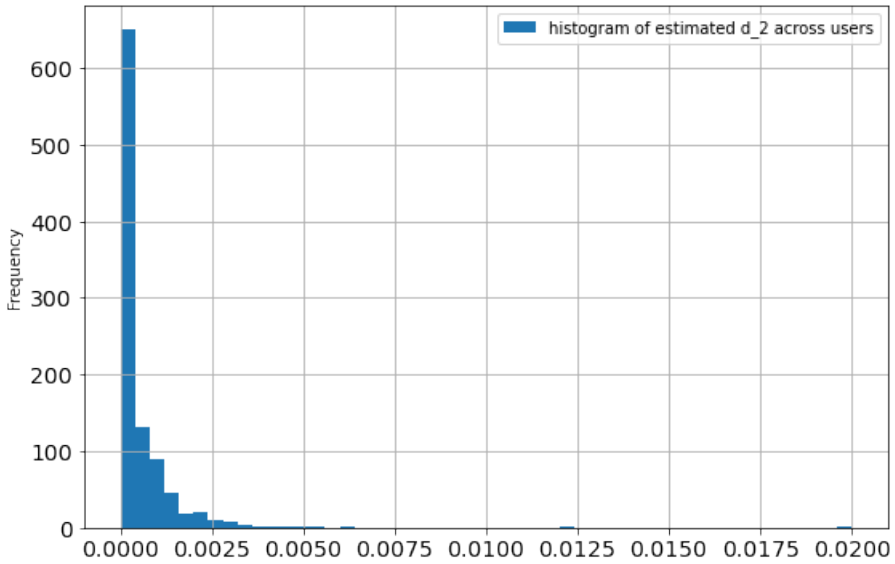

Figure 8: A histogram of the estimated values of d_2 for the 986 tracked customers in the DunnHumby dataset as described in Section 5.

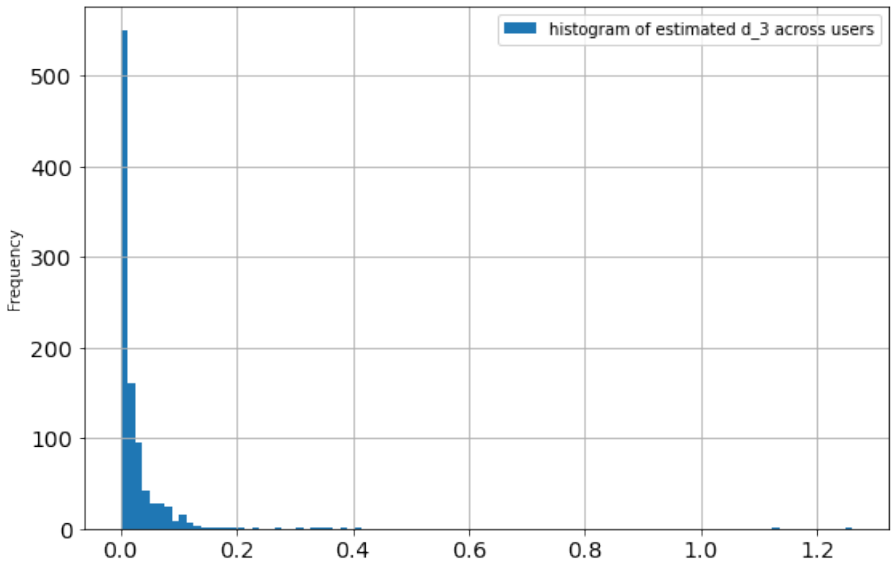

Figure 9: A histogram of the estimated values of d_3 for the 986 tracked customers in the DunnHumby dataset as described in Section 5.

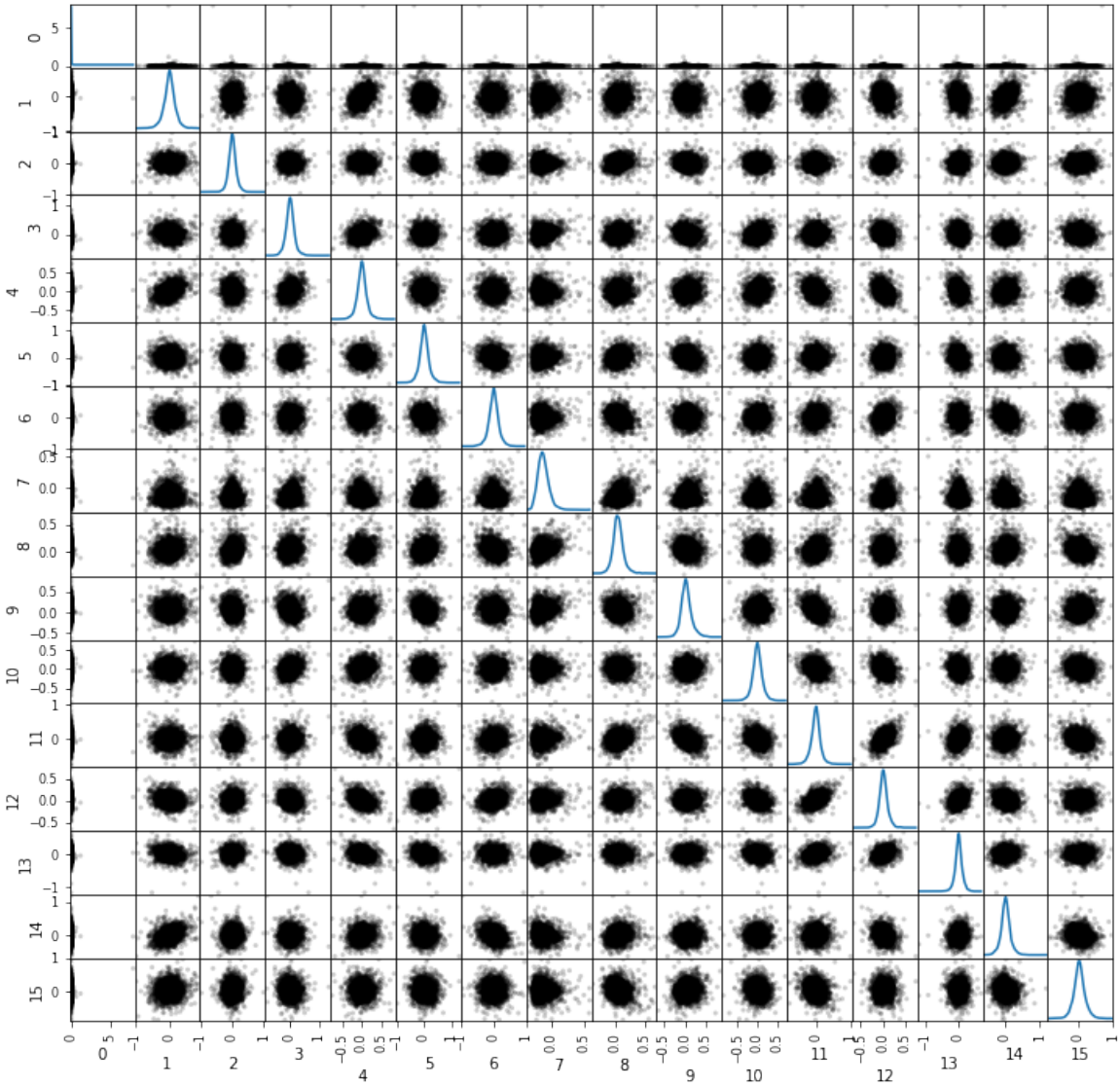

Figure 10: Goods/products: A scatter plot matrix displaying all pairwise relationships among the 4,737 estimated columns of $\boldsymbol{A}$. Each row of attribute matrix $\boldsymbol{A}$ measures exposure to one of the 16 latent attributes for which there is demand. This is an 'embedding' of goods in the DunnHumby dataset, as described in Section 5.

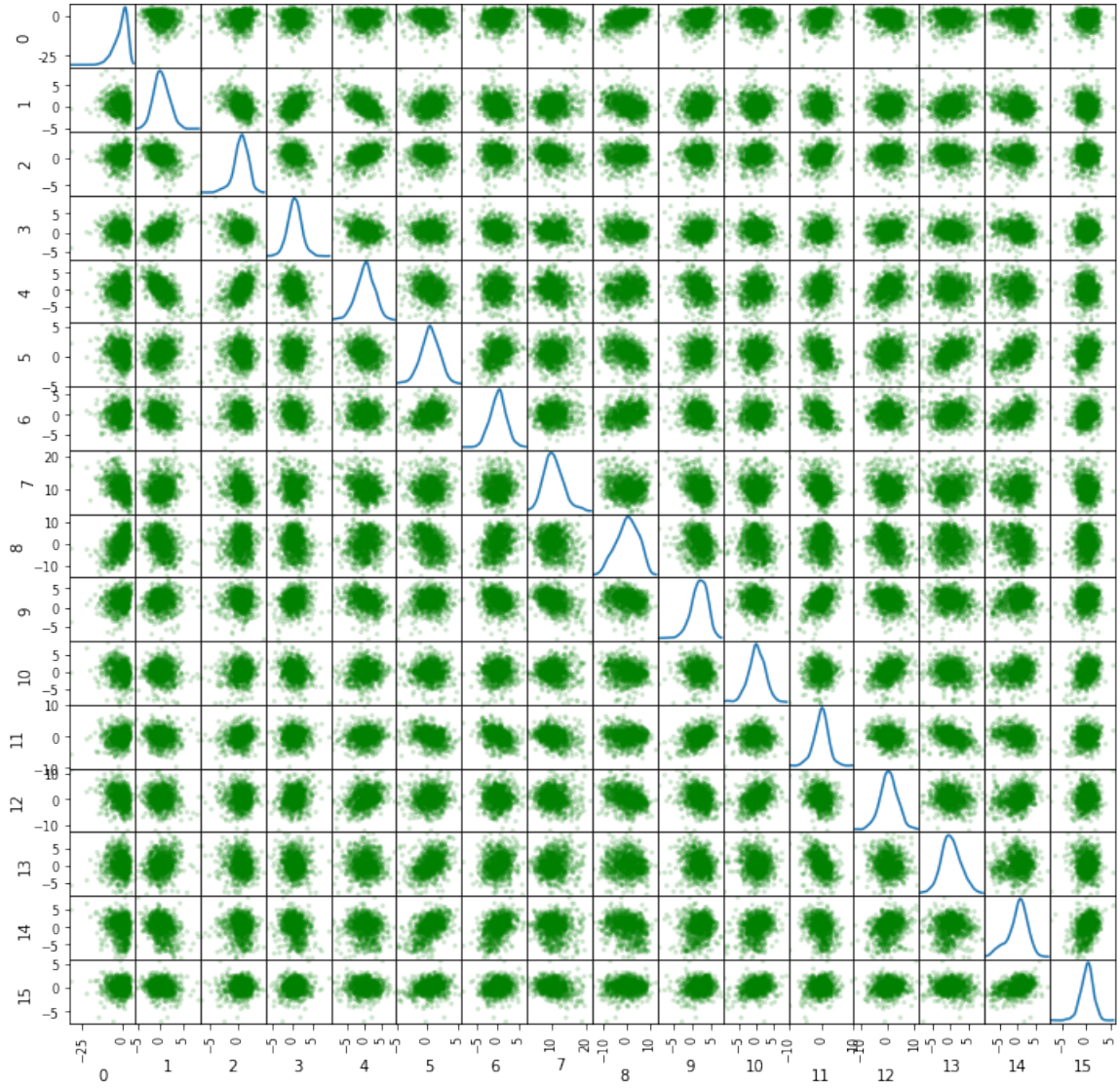

Figure 11: Consumers: A scatter plot matrix displaying all pairwise relationships among the 986 estimated rows of $\boldsymbol{b}$. Each column of the matrix $\boldsymbol{b}$ measures preferences for one of the latent attributes. This is an 'embedding' of the consumers tracked in the DunnHumby dataset, as described in Section 5.

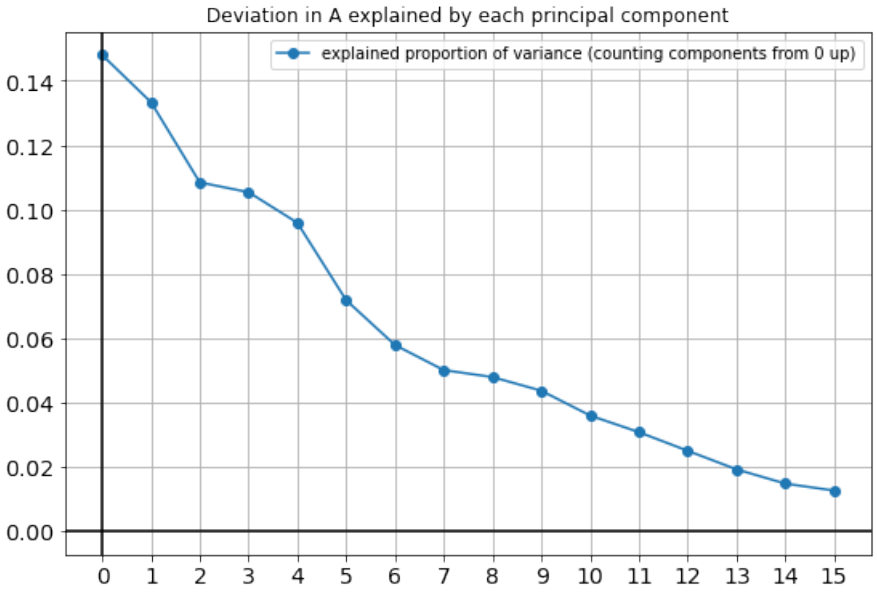

Figure 12: A plot of squared eigenvalues obtained in a principal component decomposition of the estimated matrix $\boldsymbol{A}$ as described in Section B.8. They are displayed in decreasing order.