Estimating Very Large Demand Systems

Joshua Lanier* Jeremy Large† John Quah‡

March 2022

The latest version of this paper is available here

Abstract

Whether online, in a supermarket, or elsewhere, people now assemble consumption bundles from an extremely wide variety of goods. We model this as a discrete choice between bundles, to maximize a random utility depending on the attributes of the goods in the bundle. We do not measure attributes directly, but discern them in consumption decisions. Attributes may be orders-of-magnitude fewer than goods - much reducing the effective consumption space.

Under quasi-linear preferences in our model, any given consumer tends to bundle-together goods that are price complements for her; while, on the contrary, she tends to buy substitutes separately.

We estimate consistently, at scale, by using techniques similar to negative-sampling for embedding in machine learning. This involves estimating from a big dataset every purchased good's latent attributes, jointly with every consumer's preferences over attributes.

JEL classification: C13, C34, D12, L20, L66
Keywords: discrete choice, negative sampling, demand estimation, scanner data, sparse demand
Correspondence:
Acknowledgements: Our thanks to 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 Economics Department, University of Oxford
‡Department of Economics, Johns Hopkins University

1 Introduction

We present a discrete choice random utility model for analyzing consumer demand for *large* numbers of products. The method allows for the consumer to purchase multiple units of any product and to purchase multiple products at once (one might imagine a consumer selecting a basket of goods in a supermarket). Our model allows for products to be complements or substitutes and allows for demand estimation and welfare evaluation of price changes.

In our model each product has an associated unobservable (to the analyst) vector of attributes. The consumer purchases a basket of products to maximize utility which is derived from the attributes of the products. If the dimensions of the attribute space are much smaller than the number of products we greatly reduce the effective size of the consumption space. While we gain tractability from this smaller consumption space we allow vectors of attributes to be substitutes or complements and so our model retains a considerable degree of flexibility in the types of behavior it can describe.

The parameters of our model can be estimated by maximum likelihood methods but this proves intractable for applications where the number of goods is large (note that the number of baskets which a consumer may purchase is much larger than the number of goods). This intractability motivates our introduction of a new estimation technique suitable for estimating parameters in a discrete choice model with a large choice spaces. The basic idea of the technique is to simulate random variables and then perform maximum likelihood conditioning on the simulated 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 unconditional likelihood function is intractable.

Our model can be explored via a simple example. Imagine a customer in a teashop, selecting a bundle to consume. Three goods are on offer: Tea, Biscuits, and Cake. Each, for the sake of argument, costs £1 and is offered only in unit quantity. A typical bundle, then, might be represented by a quantity-vector q, where $q = (1, 0, 1)'$ would represent the option of 'tea and cake'; while $q = (0, 0, 0)'$ would represent 'nothing'.

So the consumer selects from $2^3 = 8$ bundles. Our project is simply to study a multinomial choice model across these eight options. This idea, that these options are themselves

bundles of underlying goods, has been well developed, for example in Gentzkow (2007).

However, in the very large demand systems characteristic of a contemporary economy, there are thousands of goods, not just three, each of which is at times purchased in a variety of quantities. So, the options, far from numbering only eight, may be very numerous – indeed our model has a multinomial logit structure whose alternatives are overwhelmingly numerous and may be countably infinite in number.

	tea	biscuits	cake	Components of utility			utility	demand probability	Comments
Possible bundles:				ba	aa	d1m	ba - aa - d1m		
nothing	0	0	0	**0.0**	**0.0**	**0.0**	**0.0**	**0%**	no utility (not observed …)
pot of tea	1	0	0	**2.5**	**1.3**	**1.0**	**0.3**	**0%**	just a drink
biscuits	0	1	0	**6.2**	**2.4**	**1.0**	**2.8**	**1%**	slightly desirable
slice of cake	0	0	1	**8.8**	**2.1**	**1.0**	**5.7**	**26%**	pretty good
tea and biscuits	1	1	0	**8.7**	**2.7**	**2.0**	**4.0**	**5%**	slightly more desirable
tea and cake	1	0	1	**11.3**	**3.3**	**2.0**	**6.1**	**36%**	just the ticket
cake and biscuits	0	1	1	**15.0**	**8.7**	**2.0**	**4.3**	**7%**	thirsty work
high tea	1	1	1	**17.5**	**8.8**	**3.0**	**5.7**	**25%**	a bit much …
							logits	**100%**	
Prices:	£ 1.0	£ 1.0	£ 1.0						

Parameters:	Note: try changing the yellow cells.					
Attributes:	A			b	d1	**Interpretation:**
warm	**1.0**	**0.0**	**0.5**	**5**	**1.0**	A: the latent attributes of goods
filling	**-0.5**	**1.0**	**1.1**	**5**		b: the linear contribution of the attribute to utility
crunchy	**0.0**	**1.2**	**0.8**	**1**		d1: the disutility of £1 spent (coefficient on a linear term)

Results:						
E[bundle]	**0.66**	**0.38**	**0.94**	@	£	1.98

Figure 1: An illustrative example motivated by an individual's trip to a teashop. When three goods can be bought in unit quantity there are 8 possible bundles. Each bundle's composition and expected utility is reported. As the latter serves as the bundle's logit, it gives rise to the demand probabilities. Demand probabilities in turn determine (at the base of the figure) expected demand for each of the three goods – the expected bundle. Against a yellow background some typical values are inserted for the parameters of this model, A, b, and d_1.

In tackling this profusion, we apply a Gorman-Lancaster linear characteristics model where the characteristics, or attributes, of the goods are a matter of settled fact. However, we do not measure them, but must infer them from consumption patterns.

Let's say our tearoom is offering 'warm', 'filling' and 'crunchy' consumption with its products. Tea, Biscuits and Cake deliver differing quantities of these characteristics. We collect these quantities in a matrix, A, which in this toy example is just (3×3) but in

general of shape $(K \times L)$, where K is the number of attributes and L is the number of goods in the economy. As a settled matter, we take A to be agreed upon by all consumers.

A bundle's random utility will depend upon the attributes of the goods in the bundle. They are collected as a vector $a \in \mathbb{R}^K$. The bundle 'tea and cake' is endowed with the attribute vector $a = (A_1 + A_3)$ (adding the first column of A to the last), while 'cup of tea' delivers simply $a = A_1$. For any bundle, q, then, $a = Aq$.

In determining their own preferences, the consumer pays attention to an inner product $b'a$. In our teashop example we might set $b = (5, 5, 1)'$, so establishing weightings indicating the desirability to this consumer of, respectively, any bundle's 'warm', 'filling' and 'crunchy' attributes.

In the direction of parsimony, we allow quadratic terms in a to enter only via a's L2 norm, $a'a$. We show later that this is without loss of generality within a broad family of quadratic forms. This is because consumption behaviour only identifies the latent attributes here up to a linear transformation. Finally, the consumer pays attention to the expense of the bundle, $p'q$, applying a coefficient $d_1 > 0$ to it, which serves as a decreasing index of the affluence of the consumer.

If we consider bundle utility, U, given by:

$$U(q) = b'a - a'a - d_1 p'q, \tag{1}$$

(which we will generalise later), recalling that $a = Aq$, then we stay close to the specification in Lewbel and Nesheim (2019), which points out that this yields quasi-linear preferences in which b_k plays the role of a satiation level or bliss point for attribute k.

The specification also has similarities to the heuristic model of consumer behaviour in Ruiz, Athey, and Blei (2020), although it differs in several respects – including by being quasi-concave in a and decreasing in $p'q$.[1]

Our proposal is for independent standard Gumbel errors to be added to (1) giving a random utility. The consumer selects the bundle with highest random utility, so producing a multinomial logistic distribution of outcomes in which (1) functions as the logit.

[1]Ruiz, Athey, and Blei (2020) study a model in which each good, ℓ is endowed with both an embedding vector A_ℓ and a second K-vector known as a context vector. At the cost of quasi-concavity this can be captured here by replacing $a'a$ with $a'Da$ where D may be constructed from a square matrix of zeros, 0, and an identity matrix, I, of equal size either as $\left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$ or as $\left(\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right)$. In their generative model of the consumer's behaviour she builds a basket sequentially: after adding a good to it, she decide which, if any, to add next.

All this is illustrated in Figure 1.

Like Ruiz, Athey, and Blei (2020) we use technology and techniques from machine learning in order to estimate such a model at reasonable scale; and we could be said to engage in a `good2vec` exercise using negative sampling, which is similar to `word2vec` in natural language processing. Our code, which uses the `jax` package of Google Research, is at `www.github.com/jeremy-large/RUBE`.

In the next Section we situate our ideas in the context of the literature, before moving, in Section 3 to a formal statement of our model. Section 4 presents our comparative statics results regarding price complementarity. In Sections 5 and 6 we provide formal results regarding, respectively, identification and estimation. Section 7 develops a computationally effective estimation algorithm and simulates its properties. The algorithm is applied using real data from the market research company, DunnHumby, in Section 8.

2 Literature

2.1 Consumption bundles, and continuous consumer choice

Afriat (1967)

Barnett and Serletis (2008)

Banks, Blundell, and Lewbell (1997)

Christensen, Jorgenson, and Lau (1975)

Deaton and Muellbauer (1980)

Gallant (1981)

Gallant and Golub (1984)

Lancaster (1966)

Lewbel and Nesheim (2019) - bundle sparseness

Lewbel and Pendakur (2009)

Stone (1954)

2.2 Discrete choice in the context of random utility

Anderson, Palma, and Thisse (1995)

Ben-Akiva and Lerman (1985)

Berry, Levinsohn, and Pakes (1995, 2004)

Chesher and Santos Silva (2002)

Dubin and McFadden (1984)

Hausman and McFadden (1984)

Hausman and Wise (1978)

McFadden and Train (2000) - mixed MNL

Nevo (1998)

Rosen and Small (1979)

Ruiz, Athey, and Blei (2020) - estimates at scale, machine learning

3 The model

Let $\mathbb{N}_0$ denote the set of non-negative integers and $\mathbb{N}$ denote the set of positive integers. There are L goods (call them good 1, good 2, ..., good L) which can only be consumed in non-negative integer quantities. Let $\mathbf{q} \in \mathbb{N}_0^L$ denote a consumption bundle. We assume that each good, $\ell \in \{1, \ldots, L\}$, has an associated unobservable K-vector of attributes $\boldsymbol{\alpha}_\ell \in \mathbb{R}^K$ where $K \leq L$. Let $\mathbf{A}$ be the $K \times L$ matrix with column ℓ equal to $\boldsymbol{\alpha}_\ell$. That is,

$$\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_L]$$

Thus, $\sum_{\ell=1}^{L} \boldsymbol{\alpha}_\ell q_\ell = \mathbf{Aq}$ is the attribute vector which the consumer enjoys. We call $\mathbf{A}$ an *attribute matrix*. For attribute matrix $\mathbf{A}$ let the range of $\mathbf{A}$, denoted range($\mathbf{A}$), be defined by

$$\text{range}(\mathbf{A}) = \{\mathbf{Aq} \in \mathbb{R}^K : \mathbf{q} \in \mathbb{N}_0^L\}$$

We shall assume that the utility of the consumer is a function of the attributes of the consumption bundle purchased as well as the total amount of money spent. That is, the utility of the bundle $\mathbf{q}$ which costs $m \in \mathbb{R}_+$ can be calculated as $U(\mathbf{Aq}, m)$. Such a function U is called an attribute utility function. More specifically, a function $U : \mathcal{A} \times \mathbb{R}_+ \to \mathbb{R}$ is an *attribute utility function* provided that $\mathcal{A}$ is a convex cone in $\mathbb{R}^K$. The pair $(U, \mathbf{A})$ is called an *attributes model* when $\mathbf{A}$ is an attribute matrix and U is an attribute utility function whose domain contains range($\mathbf{A}$) $\times \mathbb{R}_+$. We shall work with attributes models which are well-behaved in the following sense.

Definition 1 *An attribute utility function* $U : \mathcal{A} \times \mathbb{R}_+ \to \mathbb{R}$ *is* well-behaved *if* $U(\mathbf{a}, m)$ *is continuously differentiable, strictly concave in* $\mathbf{a}$*, and*

$$\frac{\partial U(\mathbf{a}, m)}{\partial m} < 0, \qquad \text{for all } (\mathbf{a}, m) \in \mathcal{A} \times \mathbb{R}_+.$$

An attributes model $(U, \mathbf{A})$ *is* well-behaved *if* U *is well-behaved.*

An attribute utility function $U : \mathcal{A} \times \mathbb{R}_+ \to \mathbb{R}$ is a *Quadratic Attributes utility Function* (Qua function) if U is well-behaved and satisfies

$$U(\mathbf{a}, m) = \tilde{\mathbf{b}}' \begin{bmatrix} \mathbf{a} \\ m \end{bmatrix} - \begin{bmatrix} \mathbf{a}' & m \end{bmatrix} \tilde{\mathbf{B}} \begin{bmatrix} \mathbf{a}' \\ m \end{bmatrix}, \qquad \text{for all } (\mathbf{a}, m) \in \mathcal{A} \times \mathbb{R}_+ \tag{2}$$

for some $\tilde{\mathbf{b}} \in \mathbb{R}^{K+1}$ and $(K+1) \times (K+1)$ symmetric matrix $\tilde{\mathbf{B}}$. An attributes model $(U, \mathbf{A})$ is a *Quadratic attributes model* (Qua model) if U is a Qua function. The following proposition provides a convenient representation of the Qua function.

Proposition 1 *The attribute utility function* $U : \mathcal{A} \times \mathbb{R}_+ \to \mathbb{R}$ *is a Qua function if and only if for all* $(\mathbf{a}, m) \in \mathcal{A} \times \mathbb{R}_+$

$$U(\mathbf{a}, m) = \mathbf{b}'\mathbf{a} - \mathbf{a}'\mathbf{B}\mathbf{a} - d_1 m - d_2 m^2 - 2\tilde{\mathbf{d}}'\mathbf{a}m \tag{3}$$

where $\mathbf{b} \in \mathbb{R}^K$*,* $\tilde{\mathbf{d}} \in \mathbb{R}^K$*,* $\mathbf{B}$ *is a* $K \times K$ *symmetric positive definite matrix,* $\tilde{\mathbf{d}}'\mathbf{a} \geq 0$ *for all* $\mathbf{a} \in \mathcal{A}$*,* $d_1 > 0$*, and* $d_2 \geq 0$*.*

For convenience of estimation we shall constrain the parameters of the Qua function (specifically, we set $\mathbf{B}$ in (3) to be the identity matrix). We shall show that this restriction is made without loss in our ability to explain behavior. To do this we introduce the notion of equivalent attribute models.

Definition 2 *Two attribute models* $(U, \mathbf{A})$ *and* $(\tilde{U}, \tilde{\mathbf{A}})$ *are* equivalent *if*

$$U(\mathbf{A}\mathbf{q}, m) = \tilde{U}(\tilde{\mathbf{A}}\mathbf{q}, m), \qquad \text{for all } (\mathbf{q}, m) \in \mathbb{N}_0^L \times \mathbb{R}_+ \tag{4}$$

The reason for referring to $(U, \mathbf{A})$ and $(\tilde{U}, \tilde{\mathbf{A}})$ as equivalent when (4) holds is that neither utility function nor attribute matrix are observable and so $(U, \mathbf{A})$ and $(\tilde{U}, \tilde{\mathbf{A}})$ have the exact same implications for observables.

If it were so desired one could strengthen the notion of "equivalent attribute models" to requiring that the matrices in (4) have the same dimensions and all results in this paper

would still hold. To see that requiring the matrices in (4) to be of the same dimension is indeed a strengthening consider the following example. Let $(U, \boldsymbol{A})$ be an attribute model where $\boldsymbol{A}$ is $K \times L$ and let $\tilde{\boldsymbol{A}}$ be $(K+1) \times L$ where $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ are the same for the first K rows and row $K+1$ of $\tilde{\boldsymbol{A}}$ is a L vector of all 0s. Then, it is easy to see that $(U, \boldsymbol{A})$ and $(U, \tilde{\boldsymbol{A}})$ are equivalent but involve matrices of different dimensions.

Returning to our goal of restricting the parameters of the qua function we shall say that attribute utility function U is a *standard qua function* if $U : \mathbb{R}_+ \times \mathbb{R}^{K-1} \times \mathbb{R}_+ \to \mathbb{R}$ can be expressed as

$$U(\mathbf{a}, m) = \mathbf{b}'\mathbf{a} - \mathbf{a}'\mathbf{a} - d_1 m - d_2 m^2 - 2 d_3 a_1 m \tag{5}$$

where a_1 is the first element of $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^K$, d_1, d_2, d_3 are non-negative scalars and $d_1 > 0$. Note that a standard Qua function requires attribute 1 to take non-negative values. An attributes model $(U, \mathbf{A})$ is a *standard Qua model* if U is a standard Qua function.

Proposition 2 *A standard Qua model is a Qua model and every Qua model is equivalent to some standard Qua model.*

Proposition 2 establishes that there is no loss in our ability to explain behavior from assuming that $\mathbf{B}$ in equation (2) is the identity matrix. For this reason we focus our analysis on standard Qua models.

Instead of choosing $\mathbf{q}$ to maximize a deterministic attribute model, given by $U(\mathbf{Aq}, \mathbf{p}'\mathbf{q})$, we assume that the consumer maximizes the random utility function $\tilde{U}$ defined by

$$\tilde{U}(\mathbf{Aq}, \mathbf{p}'\mathbf{q}) = U(\mathbf{Aq}, \mathbf{p}'\mathbf{q}) + \varepsilon_{\mathbf{q}} \tag{6}$$

where $\varepsilon_{\mathbf{q}}$ is a standard Gumbel random variable and $\varepsilon_{\mathbf{q}}$, $\varepsilon_{\mathbf{q}'}$ are independent when $\mathbf{q} \neq \mathbf{q}'$. Such a model $(\tilde{U}, \mathbf{A})$ is called a *random attributes model with base utility function* U.

Let $(\tilde{U}, \mathbf{A})$ be a random attributes model. The function $f : \mathbb{N}_0^L \times \mathbb{R}_{++}^L \to [0, 1]$ is the *stochastic choice function generated by* $(\tilde{U}, \mathbf{A})$ if

$$f(\mathbf{q}|\mathbf{p}) = P\left(\tilde{U}(\mathbf{Aq}, \mathbf{p}'\mathbf{q}) \geq \sup_{\tilde{\mathbf{q}} \in \mathbb{N}_0^L} \tilde{U}(\mathbf{A}\tilde{\mathbf{q}}, \mathbf{p}'\tilde{\mathbf{q}})\right), \qquad \text{for all } \mathbf{q} \in \mathbb{N}_0^L \tag{7}$$

Intuitively, $f(\mathbf{q}|\mathbf{p})$ is the probability with which the consumer purchases $\mathbf{q}$ when prices are $\mathbf{p}$. $f(\cdot|\mathbf{p})$ is a probability mass function provided (i) the supremum in (7) is attained with probability 1 and (ii) the probability of ties are 0. The following proposition gives

a convenient expression for f when $(\tilde{U}, \mathbf{A})$ is a random attributes model with standard Qua base utility function.

Proposition 3 *Let $(\tilde{U}, \mathbf{A})$ be a random attributes model with a standard Qua base utility function U and let f be the stochastic choice function generated by $(\tilde{U}, \mathbf{A})$. Then,*

$$f(\mathbf{q}|\mathbf{p}) = \frac{\exp(U(\mathbf{Aq}, \mathbf{p}'\mathbf{q}))}{\sum_{\tilde{\mathbf{q}} \in \mathbb{N}_0^L} \exp(U(\mathbf{A}\tilde{\mathbf{q}}, \mathbf{p}'\tilde{\mathbf{q}}))} \tag{8}$$

Equation (8) should be understood as saying that (i) the probability mass function of consumption is described by (8) and (ii) the denominator in (8) is finite.

4 Comparative statics

5 Identification

We discuss the identification of U from the distribution of consumption and prices. To do this we introduce notation for the parameters of the Qua model. So, let $\boldsymbol{\theta} = [\mathbf{A}, \mathbf{b}, d_1, d_2, d_3]$ denote a list of the parameters of the standard Qua. Let Θ denote all such lists. Let $U(\cdot; \boldsymbol{\theta})$ denote the utility function defined by

$$U(\mathbf{q}, \mathbf{p}; \boldsymbol{\theta}) = U(\mathbf{Aq}, \mathbf{p}'\mathbf{q}) \tag{9}$$

where $U(\mathbf{Aq}, \mathbf{p}'\mathbf{q})$ is the standard Qua function with parameters $\mathbf{b}$, d_1, d_2, and d_3. Further, let $f(\mathbf{q}|\mathbf{p}; \boldsymbol{\theta})$ denote the stochastic choice function generated by $U(\cdot; \boldsymbol{\theta})$. For the remainder of the paper we shall use $\boldsymbol{e}_k = [0, \ldots, 0, 1, 0, \ldots, 0]$ to denote K-vector whose entries are all 0 except for entry k which is a 1. Recall that a matrix V is *orthogonal* if $V'V = VV' = I$ where I denotes the identify matrix.

Proposition 4 *Let $\boldsymbol{\theta} = [\mathbf{A}, \mathbf{b}, d_1, d_2, d_3] \in \Theta$ and $\tilde{\boldsymbol{\theta}} = [\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3] \in \Theta$ where $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ are rank K. Let $E \subseteq \mathbb{R}^L_{++}$ be a non-empty open set. The following are equivalent.*

1. $f(\mathbf{q}|\boldsymbol{p}; \boldsymbol{\theta}) = f(\mathbf{q}|\boldsymbol{p}; \tilde{\boldsymbol{\theta}})$, *for all $\mathbf{q} \in \mathbb{N}_0^L$ and all $\boldsymbol{p} \in E$.*
2. $U(\mathbf{q}, \boldsymbol{p}; \boldsymbol{\theta}) = U(\mathbf{q}, \boldsymbol{p}; \tilde{\boldsymbol{\theta}})$, *for all $\mathbf{q} \in \mathbb{N}_0^L$ and all $\boldsymbol{p} \in E$.*
3. $[\boldsymbol{A}'\boldsymbol{A}, \boldsymbol{A}'\boldsymbol{b}, d_1, d_2, d_3\boldsymbol{A}'\boldsymbol{e}_1] = [\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{A}}'\tilde{\boldsymbol{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3\tilde{\boldsymbol{A}}'\boldsymbol{e}_1]$.
4. *There is a $K \times K$ orthogonal matrix V so that*

$$[\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3\boldsymbol{A}'\boldsymbol{e}_1] = [V\tilde{\boldsymbol{A}}, V\tilde{\boldsymbol{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3\tilde{\boldsymbol{A}}'\boldsymbol{e}_1]$$

Define a function ψ by

$$\psi([\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3]) = [\boldsymbol{A}'\boldsymbol{A}, \boldsymbol{A}'\boldsymbol{b}, d_1, d_2, d_3\boldsymbol{A}'\boldsymbol{e}_1] \tag{10}$$

Proposition 4 says that two parameter vectors $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$ are empirically indistinguishable if $\psi(\boldsymbol{\theta}) = \psi(\tilde{\boldsymbol{\theta}})$.

6 Estimation

Doing MLE is practically impossible because (8) is intractable. Instead we introduce ideas which still allow the parameters to be consistently estimated. We will let $\mathbf{c}$ be a random L-vector whose probability mass function conditional on price vector $\mathbf{p}$ is $f(\cdot|\mathbf{p})$ defined by (8). We will continue to use $\mathbf{q}$ to denote a deterministic vector in $\mathbb{N}_0^L$.

The reason MLE is infeasible is because the denominator in (8) has too many terms. What we propose is to construct a random set S which contains a total of J (some finite number of) elements in $\mathbb{N}_0^L$. One of the elements in S will be the true bundle demanded while the rest will be "imposter" bundles. We may then proceed to estimate the parameters of our model by performing maximum likelihood estimation condition on S. When properly implemented, this conditional MLE turns the size of the summands which must be evaluated from an infinite series (as in (8)) to the sum of only J numbers.

6.1 Signal Functions

Definition 3 *For each $\mathbf{q} \in \mathbb{N}_0^L$ let $S(\mathbf{q})$ be a random subset of $\mathbb{N}_0^L$. S is a* signal function *if (i) each realization of $S(\mathbf{q})$ contains $\mathbf{q}$, (ii) for each $Q \subseteq \mathbb{N}_0^L$ and all $\mathbf{q}, \tilde{\mathbf{q}} \in Q$,*

$$P\Big(S(\mathbf{q}) = Q\Big) = P\Big(S(\tilde{\mathbf{q}}) = Q\Big) \tag{11}$$

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

$$P\Big(S(\mathbf{q}) \in \tilde{\mathcal{Q}}\Big) = 1$$

The support of $S(\mathbf{q})$ *is the collection $\mathcal{Q}(\boldsymbol{q})$ defined by*

$$\mathcal{Q}(\boldsymbol{q}) = \Big\{Q \subseteq \mathbb{N}_0^L : P(S(\mathbf{q}) = Q) > 0\Big\}$$

The support of S *is the collection $\mathcal{Q}$ defined by*

$$\mathcal{Q} = \bigcup_{\boldsymbol{q} \in \mathbb{N}_0^L} \mathcal{Q}(\boldsymbol{q})$$

We shall specify a specific signal function S which can be computed via a pseudo-random number generator. Now, given such an signal function we shall be interested in estimating the parameters of our model via maximum likelihood estimation conditional on prices and conditional on the signals $S(\mathbf{q})$. There are two important things to note about the signal function. First, property (i) ensures that any realization of $S(\mathbf{c})$ is a set which contains the true realization of $\mathbf{c}$ as well as some imposter bundles. Property (ii) says that when trying to distinguish between whether the true bundle passed to S is $\mathbf{q}$ or $\tilde{\mathbf{q}}$ one can infer nothing beyond what can be inferred from knowing whether $\mathbf{q}$ and $\tilde{\mathbf{q}}$ are in $S(\mathbf{q})$ or not.

Proposition 5 *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 $\boldsymbol{\rho} = \boldsymbol{p}$. We have*

$$f(\boldsymbol{q}|Q, \boldsymbol{p}) = P\Big(\mathbf{c} = \mathbf{q}\Big|\mathbf{c} \in Q, \boldsymbol{\rho} = \mathbf{p}\Big), \qquad \text{for all } Q \in \mathcal{Q} \tag{12}$$

Consequently, if the probability mass function of $\boldsymbol{c}$ conditional on $\boldsymbol{\rho}$, denoted $f(\boldsymbol{q}|\boldsymbol{p})$, satisfies (8)*, then, for all $Q \in \mathcal{Q}$,*

$$f(\boldsymbol{q}|Q, \boldsymbol{p}) = \begin{cases} \dfrac{\exp(U(\mathbf{Aq}, \mathbf{p}'\mathbf{q}))}{\sum_{\tilde{\mathbf{q}} \in Q} \exp(U(\mathbf{A}\tilde{\mathbf{q}}, \mathbf{p}'\tilde{\mathbf{q}}))}, & \text{for } \mathbf{q} \in Q \\ 0, & \text{else.} \end{cases} \tag{13}$$

Our approach will be to estimate the parameters of our model via maximum likelihood conditional on prices and the output of the signal function. We shall need to place some restrictions on the signal function.

Definition 4 *A function $g : \mathbb{N}_0^L \to \mathbb{R}$ is* grounded quadratic *if*

$$g(\mathbf{q}) = \mathbf{b}'\mathbf{q} + \mathbf{q}'\mathbf{B}\mathbf{q}, \qquad \text{for all } \mathbf{q} \in \mathbb{N}_0^L \tag{14}$$

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

Definition 5 *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 $\mathbf{q}, \tilde{\mathbf{q}} \in Q$ satisfying*

$$g(\mathbf{q}) \neq g(\tilde{\mathbf{q}}) \tag{15}$$

Definition 6 *A signal function S with support $\mathcal{Q}$ is* small *if*

1. *There exists a number $J \in \mathbb{N}$ so that $|Q| \leq J$ for each $Q \in \mathcal{Q}$.*

2. *There exists a number $J' \in \mathbb{N}$ so that*

$$||\boldsymbol{q}||_1 \leq J', \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \text{ such that } \#\mathcal{Q}(\boldsymbol{q}) > 1$$

where $\mathcal{Q}(\boldsymbol{q})$ is the support of $S(\boldsymbol{q})$.

In Definition 6 item *1.* limits the size of sets in the support of S. Item *2.* says that all bundles in $Q(\boldsymbol{q})$ cannot have many more units of consumption than are in $\boldsymbol{q}$ itself.

Our consistency result relies on S being small and distinguishing.

6.2 Consistent Estimation

Suppose we have some data on N consumption bundles purchased and the prices of goods $(\boldsymbol{c}_1, \boldsymbol{\rho}_1), (\boldsymbol{c}_2, \boldsymbol{\rho}_2), \ldots, (\boldsymbol{c}_N, \boldsymbol{\rho}_N)$. 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 place the following assumption on the data.

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

$$\boldsymbol{\rho}_n, \ \mathcal{S}_{n,1}, \ \mathcal{S}_{n,2}, \ \ldots, \ \mathcal{S}_{n,I} \tag{16}$$

are independent conditional on $\mathbf{c_n}$.

We shall assume that the price vectors satisfy the following assumption.

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* $\mathbf{c}_n$ *has conditional probability mass function* $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}^*)$ *defined by* (8)*. That is,*

$$f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}^*) = P(\mathbf{c}_n = \boldsymbol{q}|\boldsymbol{\rho}_n = \boldsymbol{p}) \tag{17}$$

Further, $\boldsymbol{A}^*$ *has rank* K*.*

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

$$\boldsymbol{E}[\mathbf{c}_n\mathbf{c}_n'], \qquad \boldsymbol{E}[\boldsymbol{\rho}_n\boldsymbol{\rho}_n'], \qquad \boldsymbol{E}[\mathbf{c}_n'\mathbf{c}_n\boldsymbol{\rho}_n], \qquad \textit{and} \qquad \boldsymbol{E}[\mathbf{c}_n'\mathbf{c}_n\boldsymbol{\rho}_n'\boldsymbol{\rho}_n] \tag{18}$$

We shall estimate $\boldsymbol{\theta}$ by maximizing the following log likelihood function.

$$\mathcal{L}_N(\boldsymbol{\theta}) = \frac{1}{NI}\sum_{n=1}^{N}\sum_{i=1}^{I} \ln\Big(f(\mathbf{c}_n|\mathcal{S}_{i,n}, \boldsymbol{\rho}_n; \boldsymbol{\theta})\Big) \tag{19}$$

where $f(\mathbf{c}_n|\mathcal{S}_{i,n}, \boldsymbol{\rho}_n; \boldsymbol{\theta})$ is defined by (13). Let $\hat{\boldsymbol{\theta}}_N \in \Theta$ satisfy

$$\hat{\boldsymbol{\theta}}_N \in \underset{\boldsymbol{\theta}\in\Theta}{} \mathcal{L}_N(\boldsymbol{\theta}) \tag{20}$$

when this argmax exists.

Theorem 1 *Suppose assumptions 1-5 hold. Let* $\hat{\boldsymbol{\theta}}_N$ *satisfy* (20) *when the argmax is non-empty. Let* ψ *be defined by* (10)*. Then,*

$$\psi(\hat{\boldsymbol{\theta}}_N) \xrightarrow{a.s.} \psi(\boldsymbol{\theta}^*) \tag{21}$$

6.3 Asymptotic Distribution

We refine Proposition 4 as follows.

Proposition 6 *Let* $\boldsymbol{\theta} = [\mathbf{A}, \mathbf{b}, d_1, d_2, d_3] \in \Theta$ *and* $\tilde{\boldsymbol{\theta}} = [\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3] \in \Theta$ *where* $\boldsymbol{A}$ *and* $\tilde{\boldsymbol{A}}$ *are rank* K *and* $d_3 \neq 0$*. Let* $E \subseteq \mathbb{R}^L_{++}$ *be a non-empty open set. Then, items 1.-4. in Proposition 4 are true if and only if either of the following are true*

5. $[\boldsymbol{A}'\boldsymbol{A}, \boldsymbol{A}'\boldsymbol{b}, \boldsymbol{A}'\boldsymbol{e}_1, d_1, d_2, d_3] = [\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{A}}'\tilde{\boldsymbol{b}}, \tilde{\boldsymbol{A}}'\boldsymbol{e}_1, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3]$.

6. *There is a* $K \times K$ *orthogonal matrix* V *so that* $V\boldsymbol{e}_1 = \boldsymbol{e}_1$ *and*

$$[\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3] = [V\tilde{\boldsymbol{A}}, V\tilde{\boldsymbol{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3]$$

Definition 7 *A parameter vector* $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3] \in \Theta$ *is in* T_1-form *if*

$$[\boldsymbol{e}_1 \ \boldsymbol{A}] = [B \ \tilde{B}] \tag{22}$$

where B *is a* $K \times K$ *lower triangle matrix and* $\tilde{B}$ *is a* $K \times (L - K + 1)$ *matrix.*

A parameter vector $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3] \in \Theta$ *is in* T_2-form *if*

$$\boldsymbol{A} = [B \ \tilde{B}] \tag{23}$$

where B *is a* $K \times K$ *lower triangle matrix and* $\tilde{B}$ *is a* $K \times (L - K)$ *matrix.*

The significance of T_1 and T_2 forms is presented in the following Proposition.

Proposition 7 *Let* ψ *be defined by* (10) *and let* $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3] \in \Theta$. *Suppose that each set of* K *distinct columns of* $[\boldsymbol{e}_1 \ \boldsymbol{A}]$ *are linearly independent. The following holds.*

1. *If* $d_3 > 0$ *then there exists a unique* $\boldsymbol{\theta}^* \in \Theta$ *in* T_1*-form so that* $\psi(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}^*)$.
2. *If* $d_3 = 0$ *then there exists a unique* $\boldsymbol{\theta}^* \in \Theta$ *in* T_2*-form so that* $\psi(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}^*)$.

[Proof of Proposition 7.] First, suppose $d_3 > 0$. Let $\boldsymbol{A}_{K-1}$ denote the $K \times K - 1$ matrix whose columns are the first $K - 1$ columns of $\boldsymbol{A}$. Let $C = [\boldsymbol{e}_1, \boldsymbol{A}_{K-1}]$. By assumption we know that C is full-rank. Thus, from Lemma 1 there is a unique orthogonal matrix V and a unique upper triangle matrix B whose diagonals are positive numbers so that $C = V'B$.

Define $\boldsymbol{A}^* = V\boldsymbol{A}$ and $\boldsymbol{b}^* = V\boldsymbol{b}$. We claim that $V\boldsymbol{e}_1 = \boldsymbol{e}_1$. Let B_1 denote the first column of B. As B is a upper triangle matrix with positive numbers on the diagonal it must be the case that $B_1 = k\boldsymbol{e}_1$ where $k > 0$. Next, we know $VC = B$ (using the fact that V is orthogonal) which implies $V\boldsymbol{e}_1 = k\boldsymbol{e}_1$. Now,

$$1 = \boldsymbol{e}_1'\boldsymbol{e}_1 = \boldsymbol{e}_1'V'V\boldsymbol{e}_1 = k^2\boldsymbol{e}_1'\boldsymbol{e}_1 = k^2$$

which shows that indeed $k = 1$ (recall that $k > 0$). So, we have shown that $V\boldsymbol{e}_1 = \boldsymbol{e}_1$.

Let $\boldsymbol{\theta}^* = [\boldsymbol{A}^*, \boldsymbol{b}^*, d_1, d_2, d_3]$. Clearly, $\boldsymbol{\theta}^* \in \Theta$, $\boldsymbol{A}^*$ is rank K, and $\boldsymbol{\theta}^*$ is in T_1-form. Thus, using the fact that $V\boldsymbol{e}_1 = \boldsymbol{e}_1$, the definition of $\boldsymbol{\theta}^*$, and Proposition 6 (specifically, we use the part of Proposition 6 which says that item *6.* implies *4.*) we see that $\psi(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}^*)$.

Next, suppose $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3] \in \Theta$ is in T_1-form and $\psi(\boldsymbol{\theta}) = \psi(\tilde{\boldsymbol{\theta}})$. We show that $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$. As $\psi(\tilde{\boldsymbol{\theta}}) = \psi(\boldsymbol{\theta})$ Proposition 6 shows that there is an orthogonal matrix $\tilde{V}$ so that $\tilde{V}\boldsymbol{e}_1 = \boldsymbol{e}_1$ and

$$[\tilde{V}\boldsymbol{A}, \tilde{V}\boldsymbol{b}, d_1, d_2, d_3] = [\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}, \tilde{d}_1, \tilde{d}_2, \tilde{d}_3] \tag{24}$$

Let $\tilde{\boldsymbol{A}}_{K-1}$ denote the first $K-1$ columns of $\tilde{\boldsymbol{A}}$. As $\tilde{\boldsymbol{\theta}}$ is in T_1-form we know that $[\boldsymbol{e}_1\ \bar{\boldsymbol{A}}]$ is an upper triangle matrix. Thus, (note that $\tilde{V}\boldsymbol{e}_1 = \boldsymbol{e}_1$ implies $\tilde{V}'\boldsymbol{e}_1 = \boldsymbol{e}_1$) we see

$$[\boldsymbol{e}_1\ \boldsymbol{A}_{K-1}] = C = V'B = \tilde{V}'[\boldsymbol{e}_1\ \bar{\boldsymbol{A}}_0] \tag{25}$$

But, Lemma 1 guarantees that the factorization in (25) is unique and so $V = \tilde{V}$. Thus, from (24) we see that $\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}}$.

Next, consider the case where $d_3 = 0$. Let $\boldsymbol{A}_K$ denote the $K \times K$ matrix whose columns are the first K columns of $\boldsymbol{A}$. Let $C = \boldsymbol{A}_K$. The proof is now a simpler version of the $d_3 > 0$ case and so we omit it.

Lemma 1 *Suppose C is some $K \times K$ matrix. There exists $K \times K$ orthogonal matrix V and $K \times K$ upper triangle matrix B where the diagonals of B are positive numbers so that*

$$C = V'B \tag{26}$$

If C is full-rank then the factorization in (26) is unique.

The existence of the factorization in (26) is well-known (it's called QR factorization). Further, its uniqueness when C is full-rank is also well-known.

Lemma 2 *Let $J \in \mathbb{N}$ where $K \leq J$ and let B_n be a sequence of $K \times J$ triangle matrices. Let $B_{n,i,j}$ denote row i column j of matrix B_n. Suppose that $B_{n,i,i} > 0$ for all n and all $i \in \{1, \ldots, K\}$. Further, suppose that there exists a $K \times J$ triangle matrix B where (i) $B_{i,i} > 0$ for all $i \in \{1, \ldots, K\}$ and (ii) $B_n'B_n \to B'B$. Then, $B_n \to B$.*

We shall show first that

$$B_{n,1,j} \to B_{1,j} \qquad \text{for all } j \in \{1, \ldots, J\} \tag{27}$$

After that we show that if, for some $\bar{i} \in \{1, \ldots, K-1\}$,

$$B_{n,i,j} \to B_{i,j}, \quad \text{for all } i \in \{1, 2, \ldots, \bar{i}\} \text{ and all } j \in \{1, \ldots, J\} \tag{28}$$

then also

$$B_{n,\bar{i}+1,j} \to B_{\bar{i}+1,j}, \quad \text{for all } j \in \{1, \ldots, J\} \tag{29}$$

This will establish the result as a consequence of mathematical induction.

So, let us first show that $B_{n,1,j} \to B_{1,j}$ for all $j \in \{1, \ldots, J\}$. To proceed, note that $B_n' B_n \to B'B$ implies that

$$\sum_{j=1}^{K} B_{n,j,k} B_{n,j,i} \longrightarrow \sum_{j=1}^{K} B_{j,k} B_{j,i}, \qquad \text{for all } k, i \in \{1, 2 \ldots, J\} \tag{30}$$

Evaluating (30) with $i = k = 1$ and using the fact that B_n and B are triangle matrices gives

$$B_{n,1,1}^2 \to B_{1,1}^2$$

Using the fact that $B_{n,1,1} > 0$ we see $B_{n,1,1} \to B_{1,1}$. Now, evaluate (30) with $k = 1$, $i \in \{2, \ldots, J\}$ and using the fact that B_n and B are triangle matrices gives

$$B_{n,1,1} B_{n,1,i} \to B_{1,1} B_{1,i}$$

Now, given that we have shown $B_{n,1,1} \to B_{1,1} > 0$ it must be the case that $B_{n,1,i} \to B_{1,i}$. Thus, we have shown (27). Next, suppose (28) holds for some $\bar{i} \in \{1, \ldots, K-1\}$. Evaluating (30) with $k = i = \bar{i} + 1$ and using the fact that B_n and B are triangle matrices we see

$$\sum_{j=1}^{\bar{i}+1} B_{n,j,\bar{i}+1}^2 \longrightarrow \sum_{j=1}^{\bar{i}+1} B_{j,\bar{i}+1}^2$$

which shows $B_{n,\bar{i}+1,\bar{i}+1} \to B_{\bar{i}+1,\bar{i}+1}$ after appealing to (28) and the fact that $B_{n,\bar{i}+1,\bar{i}+1} > 0$. Now, evaluate (30) with $k = \bar{i} + 1$, $i \in \{\bar{i} + 2, \ldots, J\}$ and using the fact that B_n and B are triangle matrices gives

$$\sum_{j=1}^{\bar{i}+1} B_{n,j,\bar{i}+1} B_{n,j,i} \longrightarrow \sum_{j=1}^{\bar{i}+1} B_{j,\bar{i}+1} B_{j,i}$$

Now, given that we have shown $B_{n,\bar{i}+1,\bar{i}+1} \to B_{\bar{i}+1,\bar{i}+1} > 0$ and we have assumed (28) it must be the case that $B_{n,\bar{i}+1,i} \to B_{\bar{i}+1,i}$. Thus, (29) is shown and the proof is complete.

Let $\mathcal{T}_1 : \Theta \to \Theta$ satisfy the following. When, $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{b}, d_1, d_2, d_3]$ is such that each set of K distinct columns of $[\boldsymbol{e}_1 \ \boldsymbol{A}]$ are linearly independent then $\mathcal{T}_1(\boldsymbol{\theta})$ returns the unique parameter vector guaranteed to exist by Proposition 7 which satisfies $\psi(\mathcal{T}_1(\boldsymbol{\theta})) = \psi(\boldsymbol{\theta})$. If it is not the case that each set of K distinct columns of $[\boldsymbol{e}_1 \ \boldsymbol{A}]$ are linearly independent then let $\mathcal{T}_1(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

Lemma 3 *Let $\boldsymbol{\theta}_n$ be a sequence in Θ, let $\boldsymbol{\theta}^* = [\boldsymbol{A}^*, \boldsymbol{b}^*, d_1^*, d_2^*, d_3^*] \in \Theta$ where $d_3 > 0$, and suppose $\psi(\boldsymbol{\theta_n}) \to \psi(\boldsymbol{\theta}^*)$. Then, $\mathcal{T}_1(\boldsymbol{\theta}_n) \to \mathcal{T}_1(\boldsymbol{\theta}^*)$.*

Let $\boldsymbol{\theta}_n = [\boldsymbol{A}_n, \boldsymbol{b}_n, d_{1,n}, d_{2,n}, d_{3,n}]$. Let $\boldsymbol{A}_{n,K-1}$ denote the first $K-1$ columns of $\boldsymbol{A}_n$. Let $C_n = [\boldsymbol{e}_1, \boldsymbol{A}_{n,K-1}]$. Let $C_n = V_n' B_n$ denote the factorization given by Lemma 1. From Lemma 7 we know

$$\mathcal{T}_1(\boldsymbol{\theta}_n) = [V_n \boldsymbol{A}_n, V_n \boldsymbol{b}_n, d_1, d_2, d_3]$$

7 Algorithm design and simulation

Theorem 1's consistency result yields an estimator got by maximizing the objective, $\mathcal{L}_N(\boldsymbol{\theta})$, in (19). Gradient descent is a suitable algorithm in principle to achieve this: assuming convexity it approaches the objective's `argmax` in convergent steps.

7.1 Computational Considerations

However, by strategically adding randomness to gradient descent, we succeed in substantially improving computational efficiency. In particular, we approximate (19) by computing it only for a subset of the observations in our dataset, and not for all. This subset is known as a *batch*. To exploit processors efficiently we set its size equal to a power of two: in practice, to $2^{10} = 1024$. Therefore, the batch contains under 1% of the dataset.

At each step in our gradient descent we redraw the batch and its signal sets. This reduces the variance of the descent and limits memory load. Also to this end, batches are not independently drawn from our dataset, but rather are simply accessed sequentially running through it. Furthermore, we record the obtained gradient and use it in immediately subsequent steps according to the `adam` algorithm.

When the whole dataset has been accessed once, we consider that an *epoch* is completed, and we begin again to pass through our data, generating fresh signal sets for each revisited batch. We use signal sets of size 100 and limit the quantity of any negatively-sampled good to be no more than six.

In advance of fitting, we broadly 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 at `www.github.com/jeremy-large/RUBE`.

Our algorithm is a form of Stochastic Gradient Descent. Its aforementioned adaptations are similar to ones used in word embedding algorithms in natural language processing. In its terms, a good analogy for the asymptotic thought experiment in Theorem 1 would be to allow batch size to grow without limit.

However, 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, and how, we rediscover the parameters of M.

7.2 Simulation Design

To this end, we set M to a model previously fitted to our scanner dataset detailed in Section 8, whose parameters we call θ^*. We then use a Metropolis Hastings algorithm to simulate baskets 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 only include goods among the most prevalent 'n' (which might, say, be 50 or 2500) goods in the DunnHumby data; and
- we limit our attention to baskets 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. Moreover, by looking only at baskets of a fixed length we may introduce bias. Nevertheless, we apply a Metropolis Hastings algorithm defined as follows:

1. Draw a basket, q_0, uniformly from $(0, 1, 2, 3, 4, 5, 6)^L$ subject to containing exactly nine non-zero entries. Set $q = q_0$.

2. From q, draw a small Signal Set of size 2. This is a set containing q, and one proposed negative sample, q_{neg}.

   Signal Set construction ensures that the proposal distribution, $q_{neg}|q$, has the property of *detailed balance*. It also preserves the number of distinct items in the basket.

3. Evaluate the utility of q_{neg}, and calculate its difference to the utility of q. The Metropolis Hastings acceptance ratio is the exponent of this difference.[2]

4. If we accept q_{neg} we set $q = q_{neg}$, otherwise 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.

7.3 Simulation Results

We run two simulations:

- a Precise Simulation, with larger than normal batch size of 8192 and smaller than normal universe of $L = 50$. Here we expect to see good performance.

- a Commensurate Simulation with a standard batch size of 1024 and a vocabulary size of $L = 2500$. 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 in an asymptotic limit theory. 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 will introduce ongoing ergodic perturbations around the truth.

The function $\psi(\cdot)$ is defined in (10). Its first term is a symmetric matrix containing pairwise inner products of the goods' embedding vectors. The second term contains the pairwise inner products of customers' embeddings with those of the goods. The third and fourth terms are scalars, d_1 and d_2, governing the disutility of expenditure. We study the timeseries of these items across epochs.

Although the simulated data has no reality, the original model, M, from which it is drawn, was fitted to real data. Some interpretation of the first 20 products of M is provided in Table 1.

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

	Product	Units	Manufacturer code
1	FLUID MILK WHITE ONLY	GA	69
2	BANANAS	LB	2
3	FLUID MILK WHITE ONLY		69
4	SHREDDED CHEESE	OZ	69
5	MAINSTREAM WHITE BREAD	OZ	69
6	SOFT DRINKS 12/18 15PK CAN CAR	OZ	1208
7	EGGS - X-LARGE	DZ	69
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	SFT DRNK 2 LITER BTL CARB INCL	LTR	69
13	MAINSTREAM WHITE BREAD	OZ	910
14	SALAD BAR FRESH FRUIT		2
15	CANDY BARS (SINGLES)	OZ	693
16	SFT DRNK 2 LITER BTL CARB INCL	LTR	1208
17	STRAWBERRIES	OZ	5937
18	SFT DRNK SNGL SRV BTL CARB (EX)	OZ	103
19	HEAD LETTUCE	CT	673
20	HOT DOG BUNS	OZ	69

Table 1: The 20 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 $(A'A)_{1,1}$, which, because it is $A_1'A_1$, is the magnitude of the embedding vector for the most common good, a milk. We also study $A_1'A_2$ which is the inner product of the vectors of the top two goods here (milk/bananas). Finally, $A_4'A_7$ will be of interest: it relates purchases of shredded cheese and of large eggs.

In simulating our model, we will pay particular interest to $(A'A)_{1,1}$, which, because it is $A_1'A_1$, is the squared magnitude of the embedding vector for the most common product, a milk. We also study $A_1'A_2$ which is the inner product of the vectors of the top two products here. Finally, $A_4'A_7$ will be of interest, because it is negative: loosely, we may say that shredded cheese and large eggs are co-present in bundles, perhaps related to their use together in standard cooking recipes such as for omelets.

7.3.1 Precise simulation: results

41,700 simulated baskets were collected, each of which drew nine items from a universe of 50 goods. It was taken as known that $K = 12$. We ran 10,000 epochs. There were 614

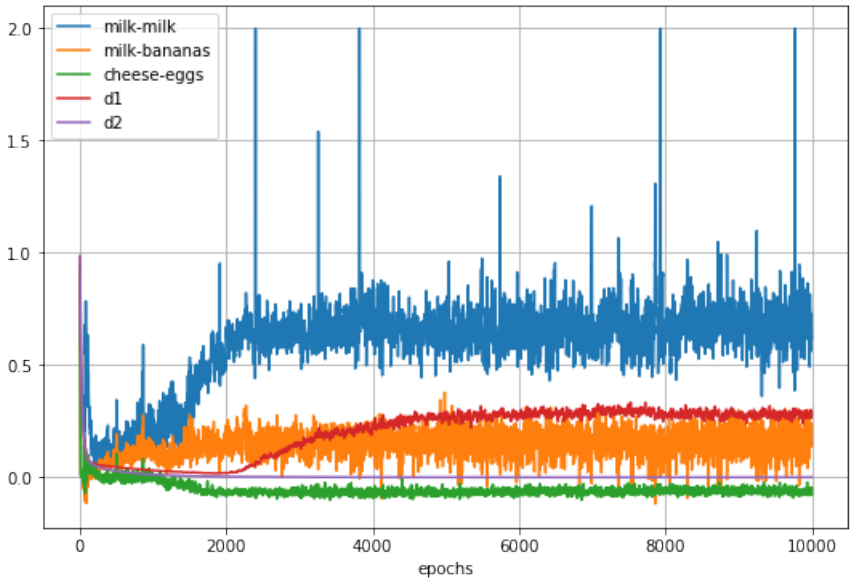

Figure 2: Precise Simulation: The evolution of various estimates as epochs pass during the estimation. Parameter estimates greater than 2 are replaced with 2. The true values are as follows: $A_1'A_1 = 0.76$ ('milk-milk') ; $A_1'A_2 = 0.15$ ('milk-bananas'); $A_4'A_7 = -0.06$ ('cheese-eggs'); $d_1 = 0.34$; $d_2 = 0.000001$.

parameters to estimate.

Figure 2 plots certain parameter values as epochs pass during estimation. Estimates fall precipitously initially and then rise to establish ergodic distributions after around 4,000 epochs.

To assess how closely these distributions are to the truth, we average fitted parameter values in the final 50 epochs, and plot these against true parameter values in Figures 3 and 4. Over all, and pending better methods of inference, the fit seems quite acceptable in this Precise Simulation.

7.3.2 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 Gorman-Lancaster linear characteristics for every one of these. It was again taken as known that $K = 12$, so that in all there were 30,014 known free parameter values to be recovered. 112,000 simulated baskets were collected, which is about half as many as we will have access to in our empirical application. Each basket again contained nine items (in various quantities). We report only the first 500 epochs.

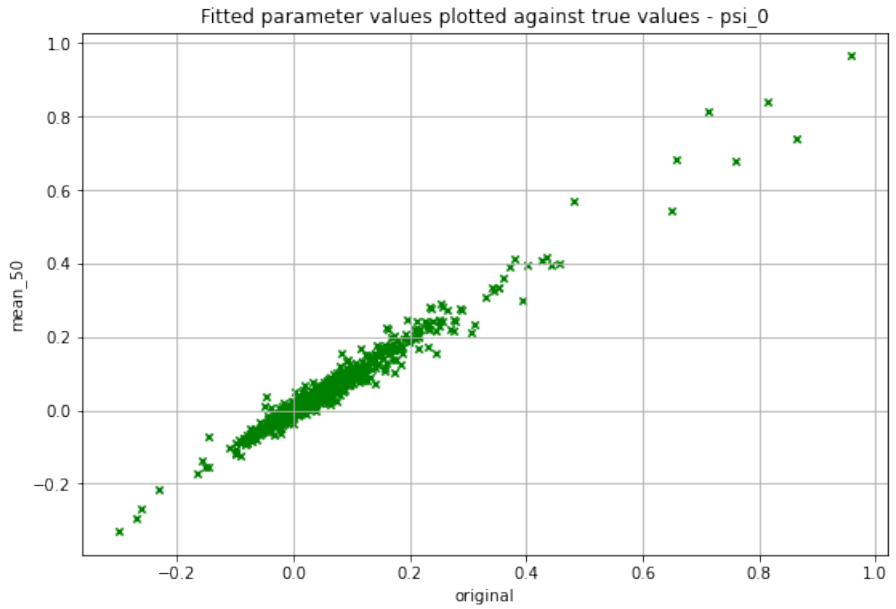

Figure 3: Precise Simulation: A cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true values. We average fitted parameter values in the final 50 epochs, and plot these against true parameter values. The 1,275 quantities plotted here are $\{A_i' A_j : 1 \leq i \leq j \leq 50\}$.

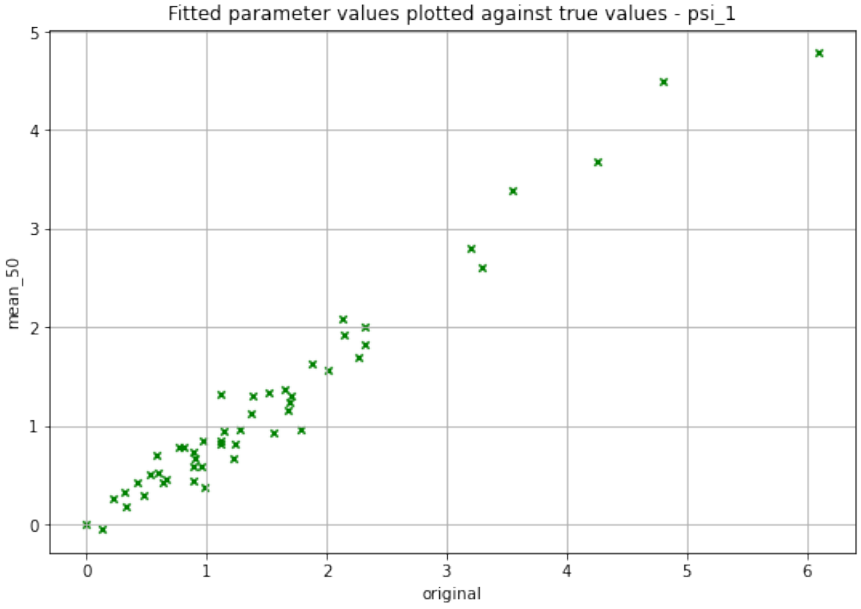

Figure 4: Precise Simulation: A cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true values. We average fitted parameter values in the final 50 epochs, and plot these against true parameter values. The 50 quantities plotted here are $\{b' A_j : 1 \leq j \leq 50\}$.

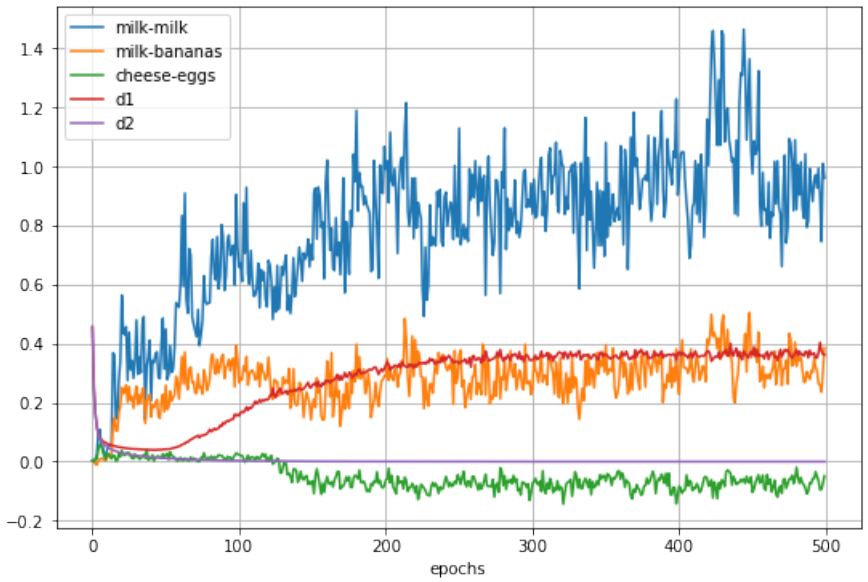

Figure 5: Commensurate Simulation: The evolution of various estimates as epochs pass during the estimation. Parameter estimates greater than 2 are replaced with 2. The true values are as follows: $A_1'A_1 = 0.76$ ('milk-milk') ; $A_1'A_2 = 0.15$ ('milk-bananas'); $A_4'A_7 = -0.06$ ('cheese-eggs'); $d_1 = 0.34$; $d_2 = 0.000001$.

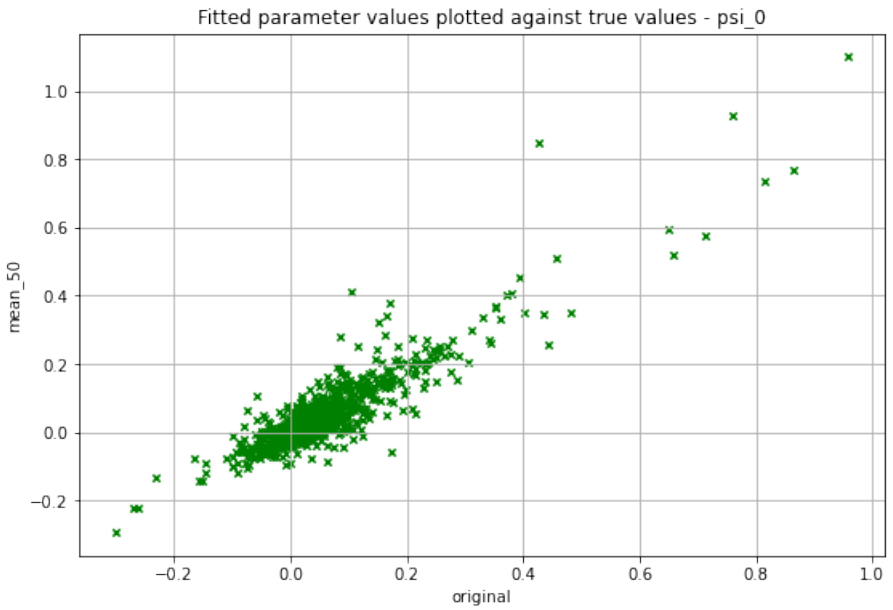

Figure 6: Commensurate Simulation: A cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true values. We average fitted parameter values in the final 50 epochs, and plot these against true parameter values. The 1,275 quantities plotted here are $\{A_i'A_j : 1 \leq i \leq j \leq 50\}$.

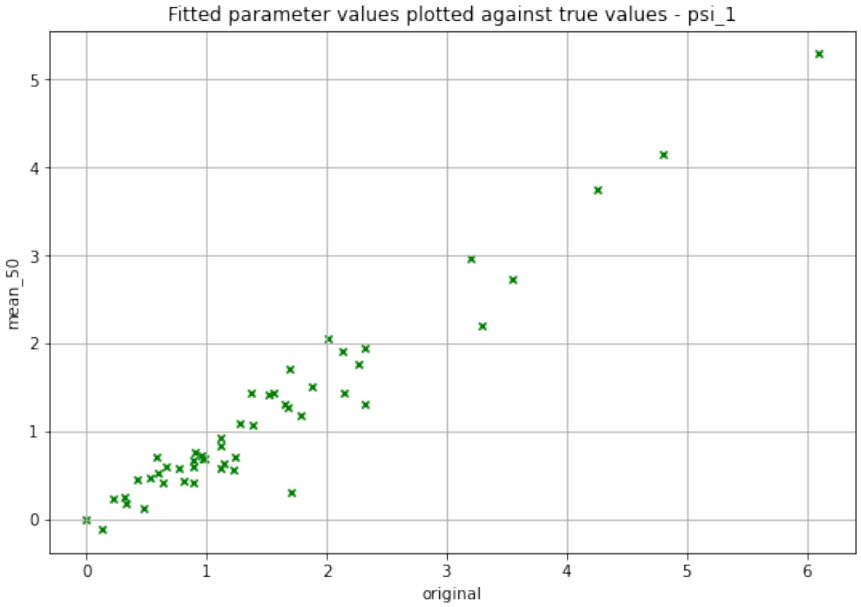

Figure 7: Commensurate Simulation: A cross-plot of some of the fitted values of $\psi(\theta^*)$ against their true values. We average fitted parameter values in the final 50 epochs, and plot these against true parameter values. Although there are estimates for 2,500 goods, the 50 quantities plotted here are $\{b'A_j : 1 \leq j \leq 50\}$.

Figure 5 plots certain parameter values as epochs pass during estimation. This is a less stable picture but by eye, it seems that after around 300 epochs estimates are settling into ergodic distributions. This is noticeably sooner than in the Precise Simulation, due presumably to the large dataset and greater variety in bundles (a 'blessing of dimensionality').

As we did with the Precise Simulation we average fitted parameter values in the final 50 epochs, and plot these against true parameter values in Figures 6 and 7. To be directly comparable to the Precise Simulation, we report information only for the set of goods which also appear in the Precise Simulation (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 8 that our real dataset will be about double the size of the one 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 (and less varied) dataset, it suggests that around 150 epochs would not be inappropriate.

8 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 are described as frequent shoppers at a retailer.

We exclude from our study the final 600,000 purchases for a future out-of-sample validation exercise. We study the remaining 2m purchases and find that they were made on 216,251 separate check-out occasions.

We view each such check-out at the till as recording a single consumption bundle. On average, bundles cost \$28.47 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.

When we distinguish products by their descriptions, manufacturers, brands and departments, as well as by their sizes bucketed into 20% bins, we find 26,633 goods. On average, each of these goods is mentioned 68 times in our data. However, we observe a long tail of rarely-purchased goods. Many goods are referred to, for example, fewer than 25 times. We track the 7,500 goods which are purchased more than 25 times. The remaining goods relate to 6% of purchases: in calculating bundle utility we account for their expense.

Some of the 2,500 households participate considerably more than others. We estimate household-specific preference parameters, $(b_u, d_{1u}, d_{2u}, d_{3u})$, for the 1,250 most active households, u. DunnHumby provides self-reported demographic information for 701 of these households. We pool all other households, u^*, such that $(b_{u^*}, d_{1u^*}, d_{2u^*}, d_{3u^*})$ is constant across them.

We use the code at `www.github.com/jeremy-large/RUBE` to estimate our model. This uses a stochastic gradient descent method. Section 7 specifies and simulates this arrangement in detail.

We hold-out a small sliver, two per cent, of our data for validation testing. Each bundle in this validation dataset is equipped 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 Base 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 7.1, this accuracy stands at 7%. But, towards the end after estimation, validation accuracy plateaus at around 35%. Over 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 set $K = 12$, so we estimate 12 latent characteristics, or embedding coordinates. The fitted $\hat{A}$ can be viewed as L observations of a multivariate random variable, whose empirical sample covariance is $\hat{A}\hat{A}'$. Applying principal component analysis to $\hat{A}$ rotates the embedding coordinates such that its empirical covariance is diagonal. The contribution of the smallest variance in this diagonalization would appear to be fairly low, measuring only 2.2% of the sum of the variances. Although this may be a reassuring indication (that there is not much work left, for our final degree of freedom to do), further work is needed to understand how best to select the tuning parameter, K, for various purposes.

As a simple sanity-check of our fit, Figure 8 compares our estimates of d_{1u}, the linear term capturing the disutility of expenditure, with self-reported household income for the 701 households where these overlap. We see that the mean estimate of d_{1u} is around 0.25 for those households reporting annual earnings in excess of \$250k, but that this rises to around 0.45 for those reporting annual earnings less than \$15k.

9 Conclusion

Embedding techniques, principally developed for text, are now also applied to household purchases - for example in Ruiz, Athey, and Blei (2020). These locate numerous retail products at points in a vector space, where proximity captures intuitive similarity. But can we interpret the embedding-space more fully?

We here present a model of the very large demand systems characteristic of many economies today. This is a Gorman-Lancaster linear characteristics specification, where all characteristics are latent. It is augmented with Gumbel errors to create a multinomial

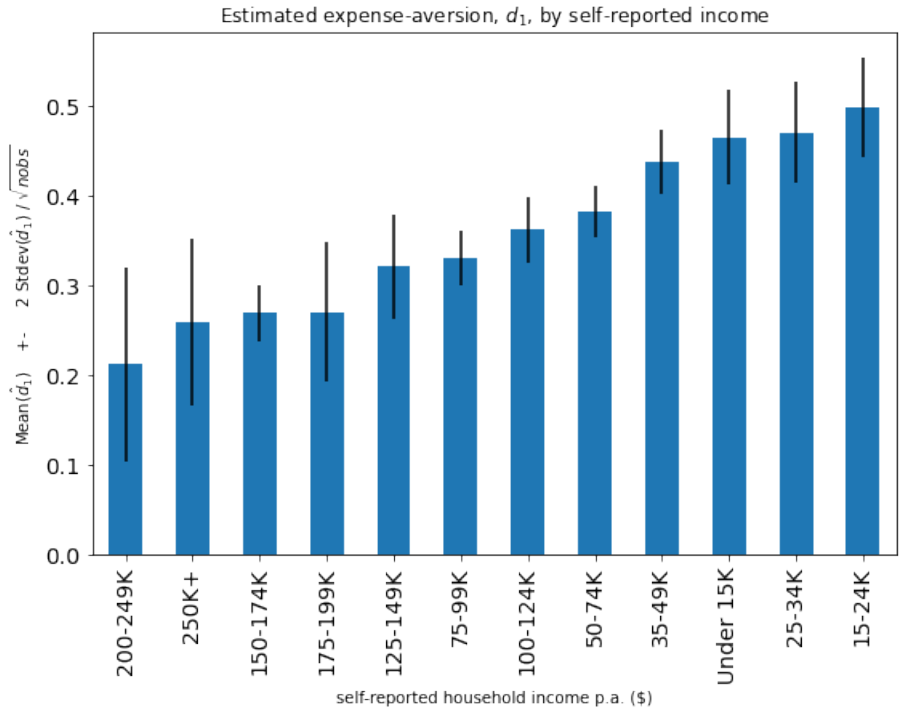

Figure 8: We estimate household-specific preference parameters, $(b_u, d_{1u}, d_{2u}, d_{3u})$, for the 1,250 most active households, u. DunnHumby provides self-reported demographic information for 701 of these households. These households are partitioned by income band, and the mean and standard deviation of d_{1u} (the linear coefficient on expenditure) in each band are displayed in this bar chart.

choice model with alternatives at least as numerous as the subsets of the set of goods on offer in the economy.

We borrow the method in machine learning known as *negative sampling* and adapt it to the current setting, generating a theory of consistent estimation. We implement this for a large dataset provided by DunnHumby and offer some observations and conclusions about our fit.

There are various avenues for future research arising from this work. There seems to be a suggestive relationship between cosine dissimilarity and price complementarity: understanding this relationship in greater detail could deepen our understanding of the embedding space. While we present a theory of consistent estimation, it would be helpful to understand theories of inference in this context, both about parameters but also about meta-parameters such as the dimensionality, K.

References

[1] S. N. Afriat. "The Construction of Utility Functions from Expenditure Data". In: *International Economic Review* 8.1 (1967), pp. 67–77. ISSN: 00206598, 14682354. URL: http://www.jstor.org/stable/2525382.

[2] Simon P. Anderson, André de Palma, and Jacques-François Thisse. "Discrete Choice Theory of Product Differentiation". In: *Journal of the Operational Research Society* 46 (1995), pp. 543–543.

[3] James. Banks, Richard Blundell, and Arthur Lewbell. "Quadratic Engel Curves and Consumer Demand". In: *Review of Economics and Statistics* 79 (1997), pp. 527–539.

[4] William A. Barnett and Apostolos Serletis. "Consumer Preferences and Demand Systems". In: *Journal of Econometrics* 147 (2008), pp. 210–224.

[5] Moshe E. Ben-Akiva and Steven R. Lerman. "Discrete Choice Analysis: Theory and Application to Travel Demand". In: 1985.

[6] Steven T. Berry, James A. Levinsohn, and Ariel Pakes. "Automobile Prices in Market Equilibrium". In: *Econometrica* 63 (1995), pp. 841–890.

[7] Steven T. Berry, James A. Levinsohn, and Ariel Pakes. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market". In: *Journal of Political Economy* 112 (2004), pp. 68–105.

[8] Andrew Chesher and João Santos Silva. "Taste variation in discrete choice models". In: *The Review of Economic Studies* 69 (2002), pp. 147–168.

[9] Laurits R. Christensen, Dale W. Jorgenson, and Lawrence J. Lau. "Transcendental Logarithmic Utility Functions". In: *American Economic Review* 65 (1975), pp. 367–383.

[10] Angus S. Deaton and John N. Muellbauer. "An Almost Ideal Demand System". In: *American Economic Review* 70 (1980), pp. 312–326.

[11] Jeffrey A. Dubin and Daniel McFadden. "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption". In: *Econometrica* 52 (1984), pp. 345–362.

[12] A. Ronald Gallant and Gene H. Golub. "Imposing curvature restrictions on flexible functional forms". In: *Journal of Econometrics* 26 (1984), pp. 295–321.

[13] Ronald A. Gallant. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form". In: *Journal of Econometrics* 15 (1981), pp. 211–245.

[14] Matthew Gentzkow. "Valuing New Goods in a Model with Complementarity: Online Newspapers". In: *American Economic Review* 97.3 (June 2007), pp. 713–744. DOI: 10.1257/aer.97.3.713. URL: https://www.aeaweb.org/articles?id=10.1257/aer.97.3.713.

[15] Jerry Hausman and Daniel McFadden. "Specification tests for the multinomial logit model". In: *Econometrica* 52 (1984), pp. 1219–1240.

[16] Jerry Hausman and David A. Wise. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences". In: *Econometrica* 46 (1978), pp. 403–427.

[17] Kelvin J. Lancaster. "A New Approach to Consumer Theory". In: *Journal of Political Economy* 74 (1966), pp. 132–157.

[18] Arthur Lewbel and Lars Nesheim. *Sparse demand systems: Corners and complements*. eng. cemmap working paper CWP45/19. London, 2019. DOI: 10.1920/wp.cem.2019.4519. URL: http://hdl.handle.net/10419/211138.

[19] Arthur Lewbel and Krishna Pendakur. "Tricks with Hicks". In: *American Economic Review* 99.3 (2009), pp. 827–863.

[20] Daniel McFadden and Kenneth E. Train. "Mixed MNL Models for Discrete Response". In: *Journal of Applied Econometrics* 15 (2000), pp. 447–470.

[21] Aviv Nevo. "Measuring Market Power in the Ready-to-Eat Cereal Industry". In: *IO: Empirical Studies of Firms & Markets* (1998).

[22] Harvey S. Rosen and Kenneth Small. "Applied Welfare Economics with Discrete Choice Models". In: 1979.

[23] Francisco J. R. Ruiz, Susan Athey, and David M. Blei. "SHOPPER: A probabilistic model of consumer choice with substitutes and complements". In: *Ann. Appl. Statist.* 14.1 (2020), pp. 1–27. ISSN: 1932-6157. DOI: 10.1214/19-AOAS1265.

[24] Richard Stone. "Linear expenditure systems and demand analysis : an application to the pattern of British demand". In: *The Economic Journal* 64 (1954), pp. 511–527.

10 Appendix

[Proof of Proposition 1.] Suppose U satisfies (2). Decompose $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{B}}$ as

$$\tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ -d_1 \end{bmatrix} \qquad \text{and} \qquad \tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{B} & \tilde{\mathbf{d}} \\ \tilde{\mathbf{d}}' & d_2 \end{bmatrix} \tag{31}$$

where $\mathbf{b} \in \mathbb{R}^K$, $d_1 \in \mathbb{R}$, $\mathbf{B}$ is a $K \times K$ symmetric matrix (as $\tilde{\mathbf{B}}$ is symmetric), $\tilde{\mathbf{d}} \in \mathbb{R}^K$, $d_2 \in \mathbb{R}$. Plugging these definitions into (2) yields (3). On the other hand, if we start by assuming that U satisfies (3) with symmetric $\mathbf{B}$ then we may define $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{B}}$ via (31) to show that U can be expressed as (2) with symmetric $\tilde{\mathbf{B}}$.

So, we must show that U, which satisfies (3) with symmetric $\mathbf{B}$, is well-behaved if and only if $\mathbf{B}$ is positive definite and

$$\tilde{\mathbf{d}}'\mathbf{a} \geq 0 \quad \text{for all } \mathbf{a} \in \mathcal{A}, \qquad d_1 > 0, \qquad \text{and} \quad d_2 \geq 0. \tag{32}$$

First, note that U is strictly concave in $\mathbf{a}$ if and only if $\mathbf{B}$ is positive definite. Next, note that if (32) holds then clearly $\partial U/\partial m < 0$. So we must show that if U is strictly decreasing in m then (32) holds.

The fact that $\mathbf{0} \in \mathcal{A}$ ensures that if $\partial U/\partial m < 0$ then $d_1 > 0$ and $d_2 \geq 0$. Next, suppose that there exists some $\tilde{\mathbf{a}} \in \mathcal{A}$ so that $\tilde{\mathbf{d}}'\tilde{\mathbf{a}} < 0$. As $\mathcal{A}$ is a cone we know that $k\tilde{\mathbf{a}} \in \mathcal{A}$ for all $k \geq 0$. Thus, we may find some large enough k so that $\partial U(k\mathbf{a}, m)/\partial m > 0$ (for some m close to 0). This establishes that $\tilde{\mathbf{d}}'\mathbf{a} \geq 0$ for all $\mathbf{a} \in \mathcal{A}$ when $\partial U/\partial m < 0$.

[Proof of Proposition 2.] First we prove that a standard Qua model is a Qua model. So, let (U, A) be a standard Qua (i.e. U satisfies (5) where d_1, d_2, d_3 are non-negative and $d_1 > 0$). U clearly satisfies (3) where $\mathbf{B}$ is the $K \times K$ identity matrix and $\tilde{\mathbf{d}} = [d_3, 0, 0, \ldots]$ (that is, $\tilde{\mathbf{d}}$ is a vector with d_3 on the first element and 0 on the remaining elements). So, by Proposition 1 we see that a standard Qua model is a Qua model.

Next, we show that every Qua model is equivalent to some standard Qua model. Let $(U, \mathbf{A})$ be a Qua model. By Proposition 1 we know that U satisfies (3) for some positive definite matrix $\mathbf{B}$ and also equation (32) hold.

As $\mathbf{B}$ is positive definite we know that both $\mathbf{B}^{1/2}$ and $\mathbf{B}^{-1/2}$ exist. Let $\mathbf{v}_1 \in \mathbb{R}^K$ be defined by

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

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

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

$$\tilde{\mathbf{d}}'\mathbf{B}^{-1/2}\mathbf{V}\mathbf{a} = d_3 a_1, \qquad \text{for all } \mathbf{a} = [a_1, a_2, \ldots, a_K] \in \mathbb{R}^K \tag{33}$$

Define an attribute utility function $\tilde{U} : \mathbb{R}_+ \times \mathbb{R}^{K-1} \times \mathbb{R}_+ \to \mathbb{R}$ by

$$\tilde{U}(\mathbf{a}, m) = \mathbf{b}'\mathbf{B}^{-1/2}\mathbf{V}\mathbf{a} - \mathbf{a}'\mathbf{a} - d_1 m - d_2 m^2 - 2d_3 a_1 m \tag{34}$$

Define an attribute matrix $\tilde{\mathbf{A}}$ by

$$\tilde{\mathbf{A}} = \mathbf{V}'\mathbf{B}^{1/2}\mathbf{A}$$

Given $d_3 \geq 0$ it is clear that $(\tilde{U}, \mathbf{A})$ is a standard Qua model. We claim that $(U, \mathbf{A})$ and

$(\tilde{U}, \tilde{\mathbf{A}})$ are equivalent. Letting $\mathbf{e}_1 = [1, 0, 0, \ldots, 0] \in \mathbb{R}^K$, $\mathbf{q} \in \mathbb{N}_0^L$, and using (33) we see

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

which shows that $(U, \mathbf{A})$ and $(\tilde{U}, \tilde{\mathbf{A}})$ are equivalent.

10.1 Proof of Proposition 3

The proof of Proposition 3 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 4 *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 5 *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 6 *Suppose* $(U, \mathbf{A})$ *is a standard Qua model. Then, for all* $\mathbf{p} \in \mathbb{R}^L_{++}$,

$$\sum_{\mathbf{q} \in \mathbf{N}_0^L} \exp\Big(U(\mathbf{A}\mathbf{q}, \mathbf{p}'\mathbf{q})\Big) < \infty \tag{35}$$

The proof of Lemma 6 appears in subsection 10.1.1. We can now prove Proposition 3. [Proof of Proposition 3.] Fix some consumption bundle $\mathbf{q} \in \mathbb{N}_0^L$ and some price vector $\mathbf{p} \in \mathbb{R}^L_{++}$. Let $\{\tilde{\mathbf{q}}^i\}_{i \in \mathbb{N}}$ be a non-repeating enumeration of the elements in $\mathbb{N}_0^L \backslash \{\mathbf{q}\}$. For $I \in \mathbb{N} \cup \{\infty\}$ let v_I be defined by

$$v_I = \ln\left(\sum_{i=1}^{I} \exp\Big(U(\mathbf{A}\tilde{\mathbf{q}}^i, \mathbf{p}'\tilde{\mathbf{q}}^i)\Big)\right) \tag{36}$$

and let $\tilde{v}_I$ be defined by

$$\tilde{v}_I = \sup_{i \leq I} \tilde{U}(\mathbf{A}\tilde{\mathbf{q}}^i, \mathbf{p}'\mathbf{q}^i) \tag{37}$$

Repeated application of Lemma 4 gives

$$\tilde{v}_I \sim \text{Gumbel}\left(v_I\right), \qquad \text{for } I \in \mathbb{N} \tag{38}$$

By Theorem 13.4(i) in billingsley86 $\tilde{v}_\infty$ 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 billingsley86 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 6 we know that $v_\infty < \infty$ and thus we see that

$$\tilde{v}_\infty \sim \text{Gumbel}(v_\infty) \tag{39}$$

Applying Lemma 5 we see

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

which completes the proof.

10.1.1 Proof of Lemma 6

The proof of Lemma 6 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 $\mathbf{v} = [v_1, \ldots, v_L] \in \mathbb{R}^L$ let $||\mathbf{v}||_1 = \sum_{\ell=1}^{L} |v_\ell|$. For $Q \in \mathbb{N}_0$, define the set $B(Q)$ by

$$B(Q) = \left\{ \mathbf{q} \in \mathbb{N}_0^L : ||\mathbf{q}||_1 = Q \right\} \tag{40}$$

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

Lemma 7 *For $Q \in \mathbb{N}_0^L$,*

$$|B(Q)| \leq (Q+1)^{L-1} \tag{41}$$

[Proof of Lemma 7.] The number $|B(Q)|$ is the number of vectors $\mathbf{q} \in \mathbb{N}_0^L$ 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 (41). We now prove Lemma 6. [Proof of Lemma 6.] Given the quadratic form of U, the fact that U is strictly concave in $\mathbf{a}$ and $\partial U/\partial m < 0$, it is clear that there is some $\mathbf{a}^* \in \mathbb{R}_+ \times \mathbb{R}^{K-1}$ and $u^* \in \mathbb{R}$ so that

$$u^* = U(\mathbf{a}^*, 0) \geq U(\mathbf{a}, m), \qquad \text{for all } \mathbf{a} \in \mathbb{R}_+ \times \mathbb{R}^{K-1} \text{ and } m \in \mathbb{R}_+ \tag{42}$$

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

$$\mathbf{p}'\mathbf{q} \leq \rho ||\mathbf{q}||_1 \tag{43}$$

Let $\boldsymbol{e}_1 = [1, 0, \ldots, 0]$. Now, applying (5), (42), the fact that $d_3 \geq 0$, and (43), we see

$$\begin{aligned} U(\mathbf{Aq}, \mathbf{p}'\mathbf{q}) &= \mathbf{b}'\mathbf{Aq} - \mathbf{q}'\mathbf{A}'\mathbf{Aq} - d_1\mathbf{p}'\mathbf{q} - d_2(\mathbf{p}'\mathbf{q})^2 - 2d_3\boldsymbol{e}_1'\boldsymbol{Aq}\mathbf{p}'\mathbf{q} \\ &\leq u^* - d_1\mathbf{p}'\mathbf{q} - d_2(\mathbf{p}'\mathbf{q})^2 - 2d_3\boldsymbol{e}_1'\boldsymbol{Aq}\mathbf{p}'\mathbf{q} \\ &\leq u^* - d_1\mathbf{p}'\mathbf{q} - d_2(\mathbf{p}'\mathbf{q})^2 \\ &\leq u^* - d_1\rho||\mathbf{q}||_1 - d_2\rho^2||\mathbf{q}||_1^2 \end{aligned} \tag{44}$$

Define $B(Q)$ by (40). Applying (44) and Lemma 7 we see

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

where the final inequality can be shown to follow using the "ratio test" for the absolute convergence of series.

10.2 Identification and Estimation Proofs

The following class of utility functions will prove useful.

Definition 8 *A utility function* $V : \mathbb{N}_0^L \times \mathbb{R}_+ \to \mathbb{R}$ *is a* quadratic expenditure modified utility function *(Quem utility function) if*

$$V(\mathbf{q}, m) = \boldsymbol{\gamma}_0'\mathbf{q} - \mathbf{q}'\mathbf{G}\mathbf{q} - \gamma_1 m - \gamma_2 m^2 - 2\boldsymbol{\gamma}_3'\mathbf{q}m \tag{45}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^L$, $\mathbf{G}$ *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* $V(\mathbf{q}, m) = 0$ *for all* $\mathbf{q} \in \mathbb{N}_0^L$ *and all* $m \in \mathbb{R}$.

Let $\boldsymbol{\gamma} = [\mathbf{G}, \boldsymbol{\gamma}_0, \gamma_1, \gamma_2, \boldsymbol{\gamma}_3]$ 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 defined by

$$V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma}) = V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) \tag{46}$$

where $V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q})$ refers to the Quem utility function, defined by (45), 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

$$h(\mathbf{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma}) = \frac{\exp(V(\mathbf{q}, \boldsymbol{p}; \boldsymbol{\gamma}))}{\sum_{\tilde{\mathbf{q}}\in Q} \exp(V(\tilde{\mathbf{q}}, \boldsymbol{p}; \boldsymbol{\gamma}))}, \qquad \text{for all } \mathbf{q} \in \mathbb{N}_0^L \tag{47}$$

where $V(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\gamma})$ is defined by (46). Also, define $h(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\gamma})$ by

$$h(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\gamma}) = h(\boldsymbol{q}|\mathbb{N}_0^L, \boldsymbol{p}; \boldsymbol{\gamma}), \qquad \text{for all } \mathbf{q} \in \mathbb{N}_0^L \tag{48}$$

where $h(\boldsymbol{q}|\mathbb{N}_0^L, \boldsymbol{p}; \boldsymbol{\gamma})$ is defined by (47). The following lemma connects the Quem utility function to the standard Qua utility function.

Lemma 8 *Let $\psi : \Theta \to \Gamma$ be defined by* (10)*. Then,*

$$U(\boldsymbol{q}, \boldsymbol{p}; \boldsymbol{\theta}) = V(\boldsymbol{q}, \boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L,\ \boldsymbol{p} \in \mathbb{R}_{++}^L,\ \boldsymbol{\theta} \in \Theta \tag{49}$$

Consequently, if $f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta})$ satisfies (8) *then*

$$f(\boldsymbol{q}|\boldsymbol{p}; \boldsymbol{\theta}) = h(\boldsymbol{q}|\boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{50}$$

Further, if $f(\cdot; Q, \boldsymbol{p}; \boldsymbol{\theta})$ satisfies (13) *for some $Q \subseteq \mathbb{N}_0^L$ then*

$$f(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\theta}) = h(\boldsymbol{q}|Q, \boldsymbol{p}; \psi(\boldsymbol{\theta})), \qquad \text{for all } \boldsymbol{q} \in \mathbb{N}_0^L \tag{51}$$

The proof is just a matter of plugging in definitions.

The following is crucial for the "identification" proof.

Lemma 9 *Let h be defined by* (47)*. 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*

$$h(\boldsymbol{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma}) \neq h(\boldsymbol{q}|Q, \boldsymbol{p}, \tilde{\boldsymbol{\gamma}}), \qquad \text{for some } Q \in \mathcal{Q},\ \boldsymbol{q} \in Q,\ \text{and all } \boldsymbol{p} \in E' \tag{52}$$

Lemma 10 *Let $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ be two full row rank $K \times L$ matrices which satisfy*

$$\boldsymbol{A}'\boldsymbol{A} = \tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}} \tag{53}$$

Define a $K \times K$ matrix V by

$$V = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\tilde{\boldsymbol{A}}' \tag{54}$$

Then, V is orthogonal and

$$V\tilde{\boldsymbol{A}} = \boldsymbol{A} \tag{55}$$

We first show that V is orthogonal. We have

$$VV' = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\boldsymbol{A}(\boldsymbol{A}\boldsymbol{A}')^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{A}(\boldsymbol{A}\boldsymbol{A}')^{-1} = I$$

which made use of (53). Next, note that

$$\begin{aligned} V &= V(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\tilde{\boldsymbol{A}}'(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{A}\tilde{\boldsymbol{A}}'(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} \\ &= \boldsymbol{A}\tilde{\boldsymbol{A}}'(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} \end{aligned} \tag{56}$$

which made use of (53). Using (56) we see

$$V'V = (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1}\tilde{\boldsymbol{A}}\boldsymbol{A}'\boldsymbol{A}\tilde{\boldsymbol{A}}'(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} = (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}'(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} = I$$

which again used (53). So, we have shown that V is orthogonal. We now show that (55) holds. So,

$$V\tilde{\boldsymbol{A}} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{A}} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}$$

which shows (55).

[Proof of Proposition 4.] It is trivial to show that *4.* $\implies$ *3.* $\implies$ *2.* $\implies$ *1.* So, we focus on the other direction. We first 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 (10). 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 9 to see that there is some non-empty open set $E' \subseteq E$ so that (52) holds. From Lemma 8 equation (49) we see that item *1.* does not hold.

Now, we show *3.* $\implies$ *4.* So, suppose *3.* holds. Define V by (54). From Lemma 10 we know that V is orthogonal and further (55). Thus, to complete the proof that *3.* $\implies$ *4.* we must show that $\boldsymbol{b} = V\tilde{\boldsymbol{b}}$. Using the fact that $\boldsymbol{A}'\boldsymbol{b} = \tilde{\boldsymbol{A}}'\tilde{\boldsymbol{b}}$ we see

$$V\tilde{\boldsymbol{b}} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\tilde{\boldsymbol{A}}'\tilde{\boldsymbol{b}} = (\boldsymbol{A}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{b} = \boldsymbol{b}$$

and so $\boldsymbol{b} = V\tilde{\boldsymbol{b}}$ as required.

[Proof of Proposition 6.] It is easily seen that *6.* implies *5.* implies *3.* (this is true even when $d_3 = 0$.) To complete the proof we show that when $d_3 \neq 0$ we have *4.* implies *6.*. So, suppose that *4.* holds and let V be the orthogonal matrix from item *4.*. Note that $d_3\boldsymbol{A}'\boldsymbol{e}_1 = \tilde{d}_3\tilde{\boldsymbol{A}}'\boldsymbol{e}_1$ implies

$$\tilde{d}_3\boldsymbol{A}'V'\boldsymbol{e}_1 = d_3\boldsymbol{A}'\boldsymbol{e}_1$$

doing left multiplication on both sides by $(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}$ gives

$$\tilde{d}_3V'\boldsymbol{e}_1 = d_3\boldsymbol{e}_1 \tag{57}$$

Using the fact that V is orthogonal we see $\tilde{d}_3^2 = d_3^2$. This implies $\tilde{d}_3 = d_3$ as $\tilde{d}_3$ and d_3 are both greater than or equal to 0. Going back to (57) and using the fact that V is orthogonal and the assumption that $d_3 > 0$ we see

$$\boldsymbol{e}_1 = V\boldsymbol{e}_1$$

Thus, we have shown that item *6.* holds.

10.2.1 Proof of Lemma 9

The proof of Lemma 9 relies on 3 lemmas.

Lemma 11 *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 $V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q})$ is a non-degenerate grounded quadratic function of $\boldsymbol{q}$.*

Let V be a Quem utility function. Suppose that $V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q})$ 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

$$V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) = 0 \qquad \text{for all } \ \boldsymbol{q} \in \mathbb{N}_0^L \ \text{ and all } \ \boldsymbol{p} \in E \tag{58}$$

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

$$\partial_{\boldsymbol{p}} V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) = -\gamma_1 \boldsymbol{q} - 2\gamma_2 \boldsymbol{p}'\boldsymbol{q}\boldsymbol{q} - 2\boldsymbol{\gamma}_3' \boldsymbol{q}\boldsymbol{q} \tag{59}$$

$$\partial^2_{\boldsymbol{p}} V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) = -2\gamma_2 \boldsymbol{q}\boldsymbol{q}' \tag{60}$$

Equation (58) implies that the expressions in both (59) and (60) are 0. From equation (60) we see that $\gamma_2 = 0$ (fix $\boldsymbol{q}$ and vary $\boldsymbol{p}$ to see this). Now, from (59) 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 $V(\boldsymbol{q}, \boldsymbol{\rho}'\boldsymbol{q})$ as

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

It is easy to show that $\boldsymbol{\gamma}_0 = 0$ and $\boldsymbol{G} = 0$ from (58). Thus, V is indeed degenerate.

Lemma 12 *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*

$$V(\boldsymbol{q}, \boldsymbol{p}'\boldsymbol{q}) \neq V(\tilde{\boldsymbol{q}}, \boldsymbol{p}'\tilde{\boldsymbol{q}}), \qquad \textit{for all } \boldsymbol{p} \in E' \tag{61}$$

Let V be a non-degenerate Quem utility function. From Lemma 11 there exists a $\bar{\boldsymbol{p}} \in E$ so that $V(\boldsymbol{q}, \bar{\boldsymbol{p}}'\boldsymbol{q})$ 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

$$V(\boldsymbol{q}, \bar{\boldsymbol{p}}'\boldsymbol{q}) \neq V(\tilde{\boldsymbol{q}}, \bar{\boldsymbol{p}}'\tilde{\boldsymbol{q}}) \tag{62}$$

But, as V is a continuous function we may find some neighborhood of $\bar{\boldsymbol{p}}$ so that (62) holds for any $\boldsymbol{p}$ in this set. This establishes (61).

Lemma 13 *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*

$$\frac{\exp(b_N)}{\sum_{n=1}^{\infty}\exp(b_n)} = \frac{\exp(\tilde{b}_N)}{\sum_{n=1}^{\infty}\exp(\tilde{b}_n)}, \qquad \textit{for each } N \in \mathbb{N} \tag{63}$$

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 (63) 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 9. [Proof of Lemma 9.] 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

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

It is easily shown that W is a non-degenerate Quem utility function. Thus, by Lemma 12 there exists $Q \in \mathcal{Q}$, $\boldsymbol{q}$, $\tilde{\boldsymbol{q}} \in Q$, and a non-empty open set $E' \subseteq E$ so that

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

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 13 establishes (52).

10.2.2 Proof of Proposition 5

[Proof of Proposition 5] Fix some $\mathbf{q} \in \mathbb{N}_0^L$ and let $Q \in \mathcal{Q}$. Consider two cases: Case 1: $\mathbf{q} \notin Q$ and Case 2: $\mathbf{q} \in Q$. Suppose Case 1 holds. As the event $\mathbf{c} \in Q$ clearly implies $\mathbf{c} \neq \mathbf{q}$ we see that the right hand side of (12) is 0. On the other hand, by property (i)

in the definition of a signal function, we see that the event $S(\mathbf{c}) = Q$ implies the event $\mathbf{c} \in Q$. But, we have already noted that the event $\mathbf{c} \in Q$ implies $\mathbf{c} \neq \mathbf{q}$ and so the left hand side of (12) is also 0. Thus, (12) holds under Case 1.

Now, consider Case 2. We write $\mathcal{S}$ for $S(\mathbf{c})$. By equation (11) and the assumption that $S(\boldsymbol{c})$ and $\boldsymbol{\rho}$ are independent conditional on $\boldsymbol{c}$

$$P\Big(\mathcal{S} = Q\Big|\mathbf{c} = \mathbf{q}, \boldsymbol{\rho}\Big) = P\Big(\mathcal{S} = Q\Big|\mathbf{c} = \tilde{\mathbf{q}}, \boldsymbol{\rho}\Big) \tag{64}$$

Now,

$$\begin{aligned}
P\Big(\mathbf{c} = \mathbf{q}\Big|\mathcal{S} = Q, \boldsymbol{\rho}\Big) &= \frac{P(\mathbf{c} = \mathbf{q} \text{ and } \mathcal{S} = Q|\boldsymbol{\rho})}{P(\mathcal{S} = Q|\boldsymbol{\rho})} \\
&= \frac{P(\mathbf{c} = \mathbf{q} \text{ and } \mathcal{S} = Q|\boldsymbol{\rho})}{\sum_{\tilde{\mathbf{q}} \in Q} P(\mathbf{c} = \tilde{\mathbf{q}} \text{ and } \mathcal{S} = Q|\boldsymbol{\rho})}, && \text{by prop (i) of def 3} \\
&= \frac{P(\mathbf{c} = \mathbf{q}|\boldsymbol{\rho}) P(S = Q|\mathbf{c} = \mathbf{q}, \boldsymbol{\rho})}{\sum_{\tilde{\mathbf{q}} \in Q} P(\mathbf{c} = \tilde{\mathbf{q}}|\boldsymbol{\rho}) P(S = Q|\mathbf{c} = \tilde{\mathbf{q}}, \boldsymbol{\rho})} \\
&= \frac{P(\mathbf{c} = \mathbf{q}|\boldsymbol{\rho})}{\sum_{\tilde{\mathbf{q}} \in Q} P(\mathbf{c} = \tilde{\mathbf{q}}|\boldsymbol{\rho})}, && \text{by (64)} \\
&= P\Big(\mathbf{c} = \mathbf{q}\Big|\mathbf{c} \in Q, \boldsymbol{\rho}\Big)
\end{aligned}$$

which establishes (12) for Case 2. Finally, (13) is a easy consequence of (12).

10.2.3 Proof of Theorem 1

The proof requires several lemmas.

Lemma 14 *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 $\bar{\delta} > 0$ such that $\psi^{-1}(B_\delta)$ is compact for all $\delta \in (0, \bar{\delta}]$.*

Lemma 15 *Suppose Assumptions 3 and 5 hold. Then,*

$$\boldsymbol{E}\left[\Big|\ln\big(h(\boldsymbol{c}_n|\mathcal{S}_{n,i}, \boldsymbol{\rho}_n; \boldsymbol{\gamma})\big)\Big|\right] < \infty, \qquad \textit{for all } \boldsymbol{\gamma} \in \Gamma, \textit{ and all } n, i \tag{65}$$

Lemma 16 *The function $\ln h(\mathbf{q}|Q, \boldsymbol{p}; \boldsymbol{\gamma})$ defined in (47) is concave in $\boldsymbol{\gamma} \in \Gamma$.*

Lemma 17 *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* $C \subseteq \Gamma$

$$\sup_{\boldsymbol{\gamma} \in C} \left| F_N(\boldsymbol{\gamma}) - F(\boldsymbol{\gamma}) \right| \xrightarrow{a.s.} 0 \tag{66}$$

Lemma 18 *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*

$$F(\bar{\boldsymbol{\gamma}}) > F(\boldsymbol{\gamma}), \qquad \textit{for all } \boldsymbol{\gamma} \neq \bar{\boldsymbol{\gamma}} \tag{67}$$

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 a* $\varepsilon > 0$ *so that*

$$F(\bar{\boldsymbol{\gamma}}) - F(\boldsymbol{\gamma}) > \varepsilon, \qquad \textit{for all } \boldsymbol{\gamma} \in C \backslash B_\delta \tag{68}$$

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 14 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 8 equation (51) we see

$$Q_N(\psi(\boldsymbol{\theta})) = \mathcal{L}_N(\boldsymbol{\theta}), \qquad \text{for all } \boldsymbol{\theta} \in \Theta \tag{69}$$

where $\mathcal{L}_N$ is defined by (19). Define $Q : \Gamma \to \mathbb{R}$ by

$$Q(\boldsymbol{\gamma}) = \boldsymbol{E} \left[\ln \Big(h(\boldsymbol{c}_1 | \mathcal{S}_{1,1}, \boldsymbol{\rho}_1; \boldsymbol{\gamma}) \Big) \right] \tag{70}$$

From Assumptions 3 and 5 and Lemma 15 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

$$Q_N(\boldsymbol{\gamma}) \xrightarrow{a.s.} Q(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \in \Gamma \tag{71}$$

From Lemma 16 it is clear that Q_N is concave and so we may use (71) and Lemma 17 to see

$$\sup_{\boldsymbol{\gamma} \in B_{\bar{\delta}}} \left| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \right| \xrightarrow{a.s.} 0 \tag{72}$$

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

$$\sup_{\boldsymbol{\gamma}\in B_{\bar{\delta}}} \Big| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \Big| \to 0 \tag{73}$$

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 (73) contain a probability 1 event (by equation (72)).

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 a $\bar{N} \in \mathbb{N}$ so that

$$||\hat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*|| < \delta, \quad \text{for all } N \geq \bar{N} \tag{74}$$

Of course (74) implies $\psi(\hat{\boldsymbol{\theta}}_N) \equiv \hat{\boldsymbol{\gamma}}_N \to \boldsymbol{\gamma}^*$ and so the proof is complete if we can show (74).

From Assumption 2, and Lemma 9, 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 18 to see that there exists an $\varepsilon > 0$ so that

$$Q(\boldsymbol{\gamma}^*) - Q(\boldsymbol{\gamma}) > \varepsilon, \qquad \text{for all } \boldsymbol{\gamma} \in B_{\bar{\delta}} \backslash B_\delta \tag{75}$$

From (73) there exists a $\bar{N} \in \mathbb{N}$ so that

$$\sup_{\boldsymbol{\gamma}\in B_{\bar{\delta}}} \Big| Q_N(\boldsymbol{\gamma}) - Q(\boldsymbol{\gamma}) \Big| < \tfrac{\varepsilon}{2}, \qquad \text{for all } N \geq \bar{N} \tag{76}$$

We claim that

$$Q_N(\boldsymbol{\gamma}^*) > Q_N(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \in B_{\bar{\delta}} \backslash B_\delta \quad \text{and all } N \geq \bar{N} \tag{77}$$

To see (77) use equations (75) and (76) 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, (77) holds. We next claim that

$$Q_N(\boldsymbol{\gamma}^*) > Q_N(\boldsymbol{\gamma}), \qquad \text{for all } \boldsymbol{\gamma} \notin B_\delta \quad \text{and all } N \geq \bar{N} \tag{78}$$

So, let $\boldsymbol{\gamma} \notin B_\delta$. If $\boldsymbol{\gamma} \in B_{\bar{\delta}}$ then (78) follows from (77) 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 (77) 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 (78) after rearranging.

B_δ is compact and so

$$\underset{\boldsymbol{\theta}\in B_\delta}{\mathcal{L}_N(\boldsymbol{\theta})} \tag{79}$$

is non-empty. Further, from (78) we see that the argmax in (20) and (79) coincide. Thus, $\hat{\boldsymbol{\theta}}_N$ is an element of the argmax in (79) (for all $N \geq \bar{N}$) and consequently $||\hat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}^*|| \leq \delta$ for all $N \geq \bar{N}$. Thus, (74) holds and so the theorem has been proved.

10.3 Proofs of lemmas used to prove Theorem 1

[Proof of Lemma 14.] 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 16.] It is clear that $V(\boldsymbol{q}, \boldsymbol{\rho}; \boldsymbol{\gamma})$, defined by (46), is linear in parameters and so it can be represented as

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

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

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

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 mcfadden74.

[Proof of Lemma 17.] 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 rockafellar70 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 (66) holds.

[Proof of Lemma 18.] 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 (68) 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 (67).

[Proof of Lemma 15.] 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

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

Using this inequality and Assumption 5 we have

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

which is what we needed to show.

10.4 Signal Function Lemma

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

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

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

Let S be a signal function. Let $\mathcal{Q}_1(\boldsymbol{q}) \equiv \mathcal{Q}(\boldsymbol{q})$ denote the support of $S(\boldsymbol{q})$. 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 19 *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

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

has full column rank then S is distinguishing.

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 $Q \in \tilde{\mathcal{Q}}$

$$\boldsymbol{b}'\boldsymbol{q} + \boldsymbol{q}'B\boldsymbol{q} = \boldsymbol{b}'\tilde{\boldsymbol{q}} + \tilde{\boldsymbol{q}}'B\tilde{\boldsymbol{q}}, \qquad \text{for all } \boldsymbol{q}, \tilde{\boldsymbol{q}} \in Q \tag{80}$$

and it is not the case that both $\boldsymbol{b} = 0$ and $B = 0$. Note that (80) 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

$$0 = \boldsymbol{b}'\boldsymbol{q}_n + \boldsymbol{q}_n'B\boldsymbol{q}_n - \sum_{m=1}^{M} \beta_m 1_{Q_m}(\boldsymbol{q}_n), \qquad \text{for all } n \in \{1, \ldots, N\} \tag{81}$$

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

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

Applying (81) we see

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

Thus, D cannot have full column rank.

Lemma 20 *If S defined by* (**??**) *is a signal function then it is distinguishing.*

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 19. To do this, let D denote the matrix in Lemma 19. 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.

$$D_2 \sim \begin{bmatrix} I & V_3^{-1}\iota \\ v_{1,1}v_{2,2}\boldsymbol{e}_1' & 1 \end{bmatrix} \sim \begin{bmatrix} I & V_3^{-1}\iota \\ 0 & 1 - \frac{v_{1,1}v_{2,2}}{v_{1,1}v_{1,2}} \end{bmatrix} = \begin{bmatrix} I & V_3^{-1}\iota \\ 0 & 1 - \frac{v_{2,2}}{v_{1,2}} \end{bmatrix} \tag{82}$$

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 19, S is distinguishing.

10.5 Derivatives and CLT

Let f satisfy (8) and let $f(\boldsymbol{q}|Q;\boldsymbol{\theta})$ satisfy (17). Let $\boldsymbol{c}$ be a random vector with pmf given by f and let $S, S_1, S_2, \ldots, S_I$ be signal functions. Let $\mathcal{S} = S(\boldsymbol{c})$ and let $\mathcal{S}_i = S_i(\boldsymbol{c})$ for each $i \in \{1, \ldots, I\}$. Let $v(\boldsymbol{q};\boldsymbol{\theta})$ be defined by

$$v(\boldsymbol{q},\boldsymbol{p};\boldsymbol{\theta}) = \frac{\partial U(\boldsymbol{q},\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{83}$$

and let $V(\boldsymbol{q},\boldsymbol{p};\boldsymbol{\theta})$ be defined by

$$V(\boldsymbol{q},\boldsymbol{p};\boldsymbol{\theta}) = \frac{\partial^2 U(\boldsymbol{q},\boldsymbol{p};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \tag{84}$$

We have the following.

Lemma 21 *The following equations are true.*

$$\begin{aligned}
\frac{\partial \ln f(\boldsymbol{c}|\mathcal{S},\boldsymbol{\rho};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= v(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta}) - \boldsymbol{E}\Big[v(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta})\Big|\mathcal{S},\boldsymbol{\rho}\Big] \\
\frac{\partial^2 \ln f(\boldsymbol{c}|\mathcal{S},\boldsymbol{\rho};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} &= V(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta}) - \boldsymbol{E}\Big[V(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta})\Big|\mathcal{S},\boldsymbol{\rho}\Big] - \mathbf{Var}\left[v(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta})|\mathcal{S}\right] \\
\mathbf{Var}\left[\frac{1}{I}\sum_{i=1}^{I}\frac{\partial \ln f(\boldsymbol{c}|\mathcal{S},\boldsymbol{\rho};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] &= \boldsymbol{E}\Big[\mathbf{Var}\left[v(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta})\big|\mathcal{S},\boldsymbol{\rho}\right]\Big] \\
&\quad - \left(\frac{I-1}{I}\right)\mathbf{Var}\left[\boldsymbol{E}\Big[\boldsymbol{E}[v(\boldsymbol{c},\boldsymbol{\rho};\boldsymbol{\theta})|\mathcal{S},\boldsymbol{\rho}]\Big|\boldsymbol{c},\boldsymbol{\rho}\Big]\right]
\end{aligned}$$

For $Q \subseteq \mathbb{N}_0^L$ let $V(Q;\boldsymbol{\theta})$ be defined by

$$V(Q;\boldsymbol{\theta}) = \sum_{\boldsymbol{q}\in Q} f(\boldsymbol{q}|Q;\boldsymbol{\theta})\frac{\partial U(\boldsymbol{q};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Define $W_1(\boldsymbol{q}|Q;\boldsymbol{\theta})$ by

$$W_1(\tilde{\boldsymbol{q}}|Q;\boldsymbol{\theta}) = \sum_{\boldsymbol{q}\in Q} f(\boldsymbol{q}|Q;\boldsymbol{\theta}) \left[\frac{\partial^2 U(\tilde{\boldsymbol{q}};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 U(\boldsymbol{q};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right]$$

and define $W_2(Q;\boldsymbol{\theta})$ by

$$W_2(Q;\boldsymbol{\theta}) = \sum_{\boldsymbol{q}\in Q} f(\boldsymbol{q}|Q;\boldsymbol{\theta}) \left(\frac{\partial U(\boldsymbol{q};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - V(Q;\boldsymbol{\theta}) \right) \left(\frac{\partial U(\boldsymbol{q};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - V(Q;\boldsymbol{\theta}) \right)'$$

It can be verified that

$$\frac{\partial \ln f(\tilde{\boldsymbol{q}}|Q;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial U(\tilde{\boldsymbol{q}};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - V(Q;\boldsymbol{\theta})$$

and that

$$\frac{\partial^2 \ln f(\tilde{\boldsymbol{q}}|Q;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = W_1(\tilde{\boldsymbol{q}}|Q;\boldsymbol{\theta}) - W_2(Q;\boldsymbol{\theta})$$