Random Projection Estimation of Discrete-Choice Models with Large Choice Sets[*]

Khai X. Chiong[†] Matthew Shum[‡]

Abstract

We introduce *random projection*, an important dimension-reduction tool from machine learning, for the estimation of aggregate discrete-choice models with high-dimensional choice sets. Initially, high-dimensional data are projected into a lower-dimensional Euclidean space using random projections. Subsequently, estimation proceeds using cyclical monotonicity moment inequalities implied by the multinomial choice model; the estimation procedure is semi-parametric and does not require explicit distributional assumptions to be made regarding the random utility errors. Our procedure is justified via the Johnson-Lindenstrauss Lemma – the pairwise distances between data points are preserved through random projections. The estimator works well in simulations and in an application to a supermarket scanner dataset.

Keywords: *discrete choice models, large choice sets, random projection, machine learning, semiparametric, cyclical monotonicity, Johnson-Lindenstrauss Lemma*

JEL: *C14, C25, C55*

1. Introduction

Discrete-choice models are a staple of empirical research in marketing and economics. Many applications of discrete-choice models in marketing, especially those utilizing supermarket scanner dataset to model household brand choice, have a challenging feature that households have a large number of options to consider. For example, supermarkets typically stock hundreds of SKUs in the cereal or soft drink categories alone, and the number of choices increases exponentially when one allows households to purchase combination

[*] First draft: February 29, 2016. This draft: April 10, 2017. We thank Serena Ng, Hiroaki Kaido, Michael Leung, Sergio Montero, Harry Paarsch, Alejandro Robinson, Frank Wolak, and seminar participants at Stanford GSB, WUSTL Olin, UT-Austin, UBC, "Machine Learning: What's in it for Economics?" (University of Chicago), the Econometric Society Australasian meetings (Sydney, July 2016), Optimal Transport and Economics (NYU, April 2016), DATALEAD (Paris, November 2015) for helpful comments.

[†] USC Dornsife INET & Department of Economics, University of Southern California. E-mail: kchiong@usc.edu

[‡] California Institute of Technology. E-mail: mshum@caltech.edu

or bundles of brands. Estimation of aggregate discrete-choice models in which consumers face high-dimensional choice sets of this sort is computationally challenging.

In this paper, we propose a new estimator that is tractable for semiparametric multinomial models with very large choice sets. Our estimator utilizes *random projection*, a powerful dimensionality-reduction technique from the machine learning literature. To our knowledge, this is the first use of random projection for the estimation of discrete-choice models. Using random projection, we can feasibly estimate high-dimensional discrete-choice models without specifying particular distributions for the random utility errors: our approach is semi-parametric.

In random projection, vectors of high-dimensionality are replaced by random low-dimensional linear combinations of the components in the original vectors. The Johnson-Lindenstrauss Lemma, the backbone of random projection techniques, justifies that with high probability, the high-dimensional vectors are embedded in a lower dimensional Euclidean space in the sense that pairwise distances and inner products among the projected-down lower-dimensional vectors are preserved.

Specifically, we are given a d-by-l data matrix, where d is the dimensionality of the choice sets. When d is very large, we encounter computational problems that render estimation difficult: estimating semiparametric discrete-choice models is already challenging, but large choice sets exacerbate the computational challenges; moreover, in extreme cases, the choice sets may be so large that typical computers will not be able to hold the data in memory (RAM) all at once for computation and manipulation.[1]

Using the idea of random projection, we propose first, in a data pre-processing step, pre-multiplying the large d-by-l data matrix by a k-by-d (with $k \ll d$) stochastic matrix, resulting in a smaller k-by-l projected-down data matrix that is more manageable. Subsequently, we estimate the discrete-choice model using the projected-down data matrix, in place of the original high-dimensional dataset. Moreover, in this step, we estimate the discrete-choice model without needing to specify the distribution of the random utility errors by using inequalities derived from *cyclical monotonicity*: – a generalization of the notion of monotonicity for vector-valued functions which always holds for random-utility discrete-choice models.[2]

A desirable and practical feature of our procedure is that the random projection matrix is sparse, so that generating and multiplying it with the large data matrix is computationally parsimonious. For instance, when the dimensionality of the choice set is $d = 5,000$, the random projection matrix consists of roughly 99% zeros (on average), and indeed only 1% of the data matrix is ever needed or sampled.

[1] For example, Ng (2015) analyzes terabytes of scanner data that required an amount of RAM that was beyond the budget of most researchers.

[2] See Rockafellar (1970), Fosgerau & De Palma (2015), Chiong, Galichon & Shum (2016).

We use the Johnson-Lindenstrauss Lemma to prove that our random projection estimator can consistently estimate the high-dimensional discrete-choice models. Intuitively, the Johnson-Lindenstrauss Lemma shows that the pairwise distances between projected-down vectors are preserved with high probability after random projection. Therefore random projection can be used as a dimensionality-reduction techniques for any estimation procedure that can be formulated in a manner that involves only the Euclidean distances between data vectors. The discrete-choice estimating inequalities based on Cyclical Monotonicity can be formulated as such.

As an application of our procedures, we estimate an aggregate model of soft drink choice in which households choose not only which soft drink product to purchase, but also the store that they shop at. In the dataset, households can choose from over 3000 (store/soft drink product) combinations, and we use random projection to reduce the number of choices to 300, one-tenth of the original number.

1.1. Related Literature

Difficulties in estimating multinomial choice models with very large choice sets were already considered in the earliest econometric papers on discrete-choice models (McFadden (1974, 1978)). There, within the special multinomial logit case, McFadden discussed simulation approaches to estimation based on sampling the choices faced by consumers; subsequently, this "sampled logit" model was implemented in Train, McFadden & Ben-Akiva (1987). This sampling approach depends crucially on the generalized extreme value distribution, and particularly on the independence of the errors across items in the large choice set.[3]

In contrast, the approach taken in this paper is semiparametric, as we avoid making specific parametric assumptions for the distribution of the errors. Our closest antecedent is Fox (2007), who uses a maximum-score approach of Manski (1975, 1985) to estimate semiparametric multinomial choice models with large choice sets but using only a subset of the choices.[4] Identification relies on a "rank-order" assumption, which is an implication of the Independence of Irrelevant Alternatives (IIA) property, and hence can be consid-

[3] See also Davis, Dingel, Monras & Morales (2016) and Keane & Wasi (2012) for other applications of sampled logit-type discrete choice models. On a related note, Gentzkow, Shapiro & Taddy (2016) use a Poisson approximation to enable parallel computation of a multinomial logit model of legislators' choices among hundreds of thousands of phrases.

[4] Fox & Bajari (2013) use this estimator for a model of the FCC spectrum auctions, and also point out another reason whereby choice sets may be high-dimensionality: specifically, when choice sets of consumers consist of *bundles* of products. The size of this combinatorial choice set is necessarily exponentially increasing in the number of products. Even though the vectors of observed market shares will be sparse, with many zeros, as long as a particular bundle does not have zero market share across all markets, it will still contain identifying information.

ered as a generalized version of IIA. It is satisfied by exchangeability of the joint error distribution.

In contrast, our cyclical monotonicity approach allows for non-exchangeable joint error distribution with arbitrary correlation between the choice-specific error terms, but requires full independence of errors with the observed covariates.[5] Particularly, our approach accommodates models with error structures in the generalized extreme value family (ie. nested logit models; which are typically non-exchangeable distributions), and we illustrate this in our empirical application below, where we consider a model of joint store and brand choice in which a nested-logit (generalized extreme value) model would typically be used.

Indeed, Fox's rank-order property and the cyclical monotonicity property used here represent two different (and non-nested) generalizations of Manski's (1975) maximum-score approach for semiparametric binary choice models to a multinomial setting. The rank-order property restricts the dependence of the utility shocks across choices (exchangeability), while cyclical monotonicity restricts the dependence of the utility shocks across different markets (or choice scenarios).[6]

All the papers mentioned here consider estimation of multinomial choice models utilizing individual-level datasets. In contrast, the estimator we propose is for an aggregate market-level setting, in which product shares are observed and aggregated over many individuals. So far, we have not considered how to tailor our estimator specifically to individual-level data.

The ideas of random projection were popularized in the Machine Learning literature on dimensionality reduction (Achlioptas (2003); Dasgupta & Gupta (2003); Li, Hastie & Church (2006); Vempala (2000)). As these papers point out, both by mathematical derivations and computational simulations, random projection allows computationally simple and low-distortion embeddings of points from high-dimensional into low-dimensional Euclidean space. However, the random projection approach will not work with all high dimensional models. The reason is that while the reduced-dimension vectors maintain the same length as the original vectors, the individual components of these lower-dimension matrices may have little relation to the components of the original vectors. Thus, models in which the components of the vectors are important would not work with random projection.

In many high-dimensional econometric models, however, only the lengths and inner prod-

[5] Besides Fox (2007), the literature on semiparametric multinomial choice models is quite small, and includes the multiple-index approach of Ichimura & Lee (1991) and Lee (1995), and a pairwise-differencing approach in Powell & Ruud (2008). These approaches do not appear to scale up easily when choice sets are large, and also are not amenable to dimension-reduction using random projection.

[6] Haile, Hortaçsu & Kosenok (2008) refer to this independence of the utility shocks across choice scenarios as an "invariance" assumption, while Goeree, Holt & Palfrey (2005) call the rank-order property a "monotonicity" or "responsiveness" condition.

ucts among the data vectors are important– this includes least-squares regression models with a fixed number of regressors but a large number of observations and, as we will see here, aggregate (market-level) multinomial choice models where consumers in each market face a large number of choices. But it will *not* work in, for instance, least squares regression models in which the number of observations are modest but the number of regressors is large – such models call for regressor selection or reduction techniques, including LASSO or principal components.[7]

Relatedly, there is a thriving literature on the interface between Big Data and Marketing. In a special issue of Marketing Science (Chintagunta, Hanssens & Hauser (2016)), we see fruitful applications of machine learning tools (such as support vector machines) in marketing problems. In those papers (Huang & Luo (2016); Jacobs, Donkers & Fok (2016); Liu, Singh & Srinivasan (2016)), machine learning tools are used for classifications or predictions. Our paper belongs to a framework called structural econometrics, where the goal is rarely prediction, but rather, we want to estimate deep parameters of consumers utility functions.

Section 2 presents our semiparametric discrete-choice modeling framework, and the moment inequalities derived from cyclical monotonicity which we will use for estimation. In section 3, we introduce random projection and show how it can be applied to the semiparametric discrete-choice context to overcome the computational difficulties with large choice sets. We also show formally that the random-projection version of our estimator converges to the full-sample estimator as the dimension of the projection increases. Section 4 contains results from simulation examples, demonstrating that random projection works well in practice, even when choice sets are only moderately large. In section 5, we estimate a model of households' joint decisions of store and brand choice, using store-level scanner data. Section 6 concludes.

2. Modeling Framework

We consider a semiparametric multinomial choice framework. The choice-specific utilities are assumed to take a single index form, but the distribution of utility shocks is unspecified and treated as a nuisance element.[8] Specifically, an agent chooses from among $\mathcal{C} = [1, \ldots, d]$ alternatives or choices. High-dimensionality here refers to a large value of d.

[7] See Belloni, Chen, Chernozhukov & Hansen (2012), Belloni, Chernozhukov & Hansen (2014), and Gillen, Montero, Moon & Shum (2015). Neither LASSO nor principal components do not maintain lengths and inner products of the data vectors; typically, they will result in reduced-dimension vectors with length strictly smaller than the original vectors.

[8] Existing papers on semiparametric multinomial choices use similar setups (Fox (2007), Ichimura & Lee (1991), Lee (1995), Powell & Ruud (2008)).

The utility that agent i derives from choice j is $\boldsymbol{X}_j\boldsymbol{\beta}+\epsilon_{ij}$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_b)' \in \mathbb{R}^b$ are unknown parameters, and $\boldsymbol{X}_j$ is a $1 \times b$ vector of covariates specific to choice j. Now ϵ_{ij} is the error term, encompassing unobserved individual tastes and product heterogeneity. This framework is the familiar BLP model (Berry, Levinsohn & Pakes (1995)) which aggregates out individual choices so that we consider only the estimation of the mean utility $u_j \equiv \boldsymbol{X}_j\boldsymbol{\beta}$, $j = 1, \dots, d$, that agents derive from different choices. That is, we do not require or make use of individual-level characteristics, but rather product or choice-level characteristics.

Let $\boldsymbol{u} = (u_j)_{j=1}^d$, which we assume to lie in the set $\mathcal{U} \subseteq \mathbb{R}^d$. For a given $\boldsymbol{u} \in \mathcal{U}$, the probability that choice j is chosen in the aggregate is $p_j(\boldsymbol{u}) = \Pr(u_j + \epsilon_{ij} \geq \max_{k \neq j}\{u_k + \epsilon_{ik}\})$. Denote the vector of choice probabilities as $\boldsymbol{p}(\boldsymbol{u}) = (p_j(\boldsymbol{u}))_{j=1}^d$. Now observe that the choice probabilities vector $\boldsymbol{p}$ is a vector-valued function such that $\boldsymbol{p} : \mathcal{U} \to \mathbb{R}^d$.

In this paper, we assume that the utility shocks $\boldsymbol{\epsilon}_i \equiv (\epsilon_{i1}, \dots, \epsilon_{id})'$ are identically and distributed independently of $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \dots, \boldsymbol{X}_d)$, but otherwise we allow it to follow an unknown joint distribution that can be arbitrarily correlated among different choices j. We further suppress the subscript i in $\boldsymbol{\epsilon}_i$ since $\boldsymbol{\epsilon}_i$ is i.i.d across i. This leads to the following proposition:

Proposition 1. *Let $\boldsymbol{\epsilon}$ be independent of $\boldsymbol{X}$. Then the choice probability function $\boldsymbol{p} : \mathcal{U} \to \mathbb{R}^d$ satisfies* **cyclical monotonicity** *(or cyclically monotone increasing).*

Definition 1 (Cyclical Monotonicity)**:** Consider a function $\boldsymbol{p} : \mathcal{U} \to \mathbb{R}^d$, where $\mathcal{U} \subseteq \mathbb{R}^d$. Take a length L-cycle of points in $\mathcal{U}$, denoted as the sequence $(\boldsymbol{u}^1, \boldsymbol{u}^2, \dots, \boldsymbol{u}^L, \boldsymbol{u}^1)$. The function $\boldsymbol{p}$ is cyclical monotone with respect to the cycle $(\boldsymbol{u}^1, \boldsymbol{u}^2, \dots, \boldsymbol{u}^L, \boldsymbol{u}^1)$ if and only if

$$\sum_{l=1}^{L} (\boldsymbol{u}^{l+1} - \boldsymbol{u}^l) \cdot \boldsymbol{p}(\boldsymbol{u}^l) \leq 0 \tag{1}$$

where $\boldsymbol{u}^{L+1} = \boldsymbol{u}^1$. The function $\boldsymbol{p}$ is cyclical monotone on $\mathcal{U}$ if and only if it is cyclical monotone with respect to all possible cycles of all lengths on its domain (see Rockafellar (1970)). ■

Proposition 1 arises from the underlying convexity properties of the discrete-choice problem, see for example, Chiong et al. (2016); Fosgerau & De Palma (2015); Fosgerau & McFadden (2012); Shi, Shum & Song (2016). To show Proposition 1, the independence of $\boldsymbol{\epsilon}$ and $\boldsymbol{X}$ implies that the *social surplus function* of the discrete choice model, defined as,

$$\mathcal{G}(\boldsymbol{u}) = \mathbb{E}\left[\max_{j \in \{1,\dots,d\}} (u_j + \epsilon_j) \,\middle|\, \boldsymbol{u}\right]$$

is convex in $\boldsymbol{u}$. Subsequently, for each vector of utilities $\boldsymbol{u} \in \mathcal{U}$, the corresponding vector of choice probabilities $\boldsymbol{p}(\boldsymbol{u})$, lies in the subgradient of $\mathcal{G}$ at $\boldsymbol{u}$;[9] that is:

$$\boldsymbol{p}(\boldsymbol{u}) \in \partial\mathcal{G}(\boldsymbol{u}) \tag{2}$$

By a fundamental result in convex analysis (Rockafellar (1970), Theorem 23.5), the subgradient of a convex function satisfies cyclical monotonicity, and hence satisfies the CM-inequalities in (1) above. (In fact, any function that satisfies cyclical monotonicity must be a subgradient of some convex function.)

It is known that in the one-dimensional case, i.e. when $p : \mathbb{R} \to \mathbb{R}$ is a scalar function, cyclical monotonicity reduces exactly to the condition that: if $y \geq x$ then $p(y) \geq p(x)$.[10] *Therefore, just as the gradient (slope) of a scalar-valued convex function is monotone increasing, the gradient of a vector-valued convex function is cyclically monotone increasing.*

It turns out that we can rewrite cyclical monotonicity in a lesser known form which allows us to combine cyclical monotonicity with random projection later on. Moreover this form of cyclical monotonicity shows intuitively why it can be seen as a vector generalization of the scalar notion of monotonicity.

Definition 2 (Cyclical Monotonicity in terms of Euclidean norms)**:** Consider a function $\boldsymbol{p} : \mathcal{U} \to \mathbb{R}^d$, where $\mathcal{U} \subseteq \mathbb{R}^d$. Take a length L-cycle of points in $\mathcal{U}$, denoted as the sequence $(\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^L, \boldsymbol{u}^1)$. The function $\boldsymbol{p}$ is cyclical monotone with respect to the cycle $\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^L, \boldsymbol{u}^1$ if and only if

$$\sum_{l=2}^{L+1} \left(\|\boldsymbol{u}^l - \boldsymbol{p}^l\|^2 - \|\boldsymbol{u}^l - \boldsymbol{p}^{l-1}\|^2 \right) \leq 0 \tag{3}$$

where $\boldsymbol{u}_{L+1} = \boldsymbol{u}_1$, and $\boldsymbol{p}^l$ denotes $\boldsymbol{p}(\boldsymbol{u}^l)$. The function $\boldsymbol{p}$ is cyclical monotone on $\mathcal{U}$ if and only if it is cyclical monotone with respect to all possible cycles of all lengths on its domain. ■

In the above display, $\|\mathbf{a}\|^2$ denotes the *squared Euclidean* norm of a vector $\mathbf{a}$, defined as $\|\mathbf{a}\|^2 = \sum_i a_i^2$. Definitions (1) and (2) are equivalent in the sense that a function satisfies the inequalities in (1) if and only if it satisfies the inequalities in (3). This equivalence is mentioned in Villani (2003), and we provide a proof in the appendix.

[9] See Theorem 1(i) in Chiong et al. (2016). This is the Williams-Daly-Zachary Theorem (cf. McFadden (1981)), generalized to the case when the social surplus function may be non-differentiable, corresponding to cases where the utility shocks $\boldsymbol{\epsilon}$ have bounded support or follow a discrete distribution.

[10] For instance, consider the cycle (x, y, x), we then have $(y-x)p(x)+(x-y)p(y) = (y-x)(p(x)-p(y)) \leq 0$, which implies that if $y \geq x$ then $p(y) \geq p(x)$.

To gain some intuition for cyclical monotonicity, consider the simplest setup in which both u and p are scalar real numbers. In this case, the inequalities in (3) imply that u^l be closer to p^l than it is to $p^{l'}$ for any other $l' \neq l$; the monotonicity of p in u implies that the two quantities "comove" the most. In higher dimensions, when $\boldsymbol{u}$ and $\boldsymbol{p}$ are vectors, the inequalities in (3) similarly imply that $\boldsymbol{p}$ and $\boldsymbol{u}$ comove the most, in the sense that $\boldsymbol{u}^l$ be closer to $\boldsymbol{p}^l$ (in terms of the squared Euclidean distance) than it is to $\boldsymbol{p}^{l'}$ for any other $l' \neq l$.

2.1. Inequalities for Estimation

We use the cyclical monotonic inequalities in (1) to estimate the parameters $\boldsymbol{\beta}$.[11] Suppose we observe the aggregate behavior of many independent agents across M different markets. In this paper, we assume the researcher has access to such aggregate data, in which the market-level choice probabilities (or market shares) are observed. Such data structures are commonplace in aggregate demand models in empirical industrial organization (for instance, Berry & Haile (2014), Berry et al. (1995)).

Our dataset consists of $\mathcal{D} = \big((\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(M)}, \boldsymbol{p}^{(M)})\big)$, $\boldsymbol{p}^{(m)}$ denotes the $d \times 1$ vector of choice probabilities, or market shares, in market m, and $\boldsymbol{X}^{(m)}$ is the $d \times b$ matrix of covariates for market m (where row j of $\boldsymbol{X}^{(m)}$ corresponds to $\boldsymbol{X}_j^{(m)}$, the vector of covariates specific to choice j in market m). Assuming that the distribution of the utility shock vectors $\big(\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(M)}\big)$ is i.i.d. across all markets, then by Proposition 1, the cyclical monotonicity inequalities (1) will be satisfied across all cycles in the data $\mathcal{D}$: that is,

$$\sum_{l=1}^{L} (\boldsymbol{X}^{(a_{l+1})}\boldsymbol{\beta} - \boldsymbol{X}^{(a_l)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(a_l)} \leq 0, \quad \text{for all cycles } (a_l)_{l=1}^{L+1} \text{ in data } \mathcal{D}, L \geq 2 \tag{4}$$

Recall that a cycle in data $\mathcal{D}$ is a sequence of distinct integers $(a_l)_{l=1}^{L+1}$, where $a_{L+1} = a_1$, and each integer is smaller than or equal n, the number of markets.

From the cyclical monotonicity inequalities in (4), we define a criterion function which we will optimize to obtain an estimator of $\boldsymbol{\beta}$. This criterion function is the sum of squared violations of the cyclical monotonicity inequalities:

$$Q(\boldsymbol{\beta}) = \sum_{\text{all cycles in data } \mathcal{D}; L \geq 2} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})}\boldsymbol{\beta} - \boldsymbol{X}^{(a_l)}\boldsymbol{\beta} \right) \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 \tag{5}$$

[11] See also Melo, Pogorelskiy & Shum (2015) for an application of cyclical monotonicity for testing game-theoretic models of stochastic choice.

where $[x]_+ = \max\{x, 0\}$. Our estimator is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^b:\|\boldsymbol{\beta}\|=1}{\text{argmin}} \; Q(\boldsymbol{\beta}).$$

Following the tradition in semiparametric models, we need to impose a scale normalization since the scale of the utilities is not identified. The parameters are normalized such that the vector $\hat{\boldsymbol{\beta}}$ has a Euclidean length of 1. This is a standard normalization that is also used in the Maximum Rank Correlation estimator (Han (1987), Hausman, Abrevaya & Scott-Morton (1998)) and in the Maximum Score estimator (Manski (1975)). We will also impose a location normalization by setting the utilities associated with the outside good to be zero for all markets.

2.2. Example

Consider an example with three markets to illustrate our estimation approach. The data matrix is $\{\boldsymbol{p}^{(1)}, \boldsymbol{X}^{(1)}, \boldsymbol{p}^{(2)}, \boldsymbol{X}^{(2)}, \boldsymbol{p}^{(3)}, \boldsymbol{X}^{(3)}\}$. Let $\boldsymbol{u}^{(m)} = \boldsymbol{X}^{(m)}\boldsymbol{\beta}$.

First we enumerate all possible cycles (without loss of generality, we can consider only circular permutations)[12] that can be formed using three markets: $\mathcal{C} = \{(1,2,1), (1,3,1), (2,3,2), (1,2,3,1), (1,3,2,1)\}$. Then we want to find all $\boldsymbol{\beta}$ that satisfies all the inequalities below (one inequality for each distinct cycle). Our estimator works by minimizing the criterion function in Equation 5, which is equivalent to finding all $\boldsymbol{\beta}$ that minimizes the sum of squared violations of cyclical monotonicity inequalities.

$$\begin{aligned}
(1,2,1):&\quad (\boldsymbol{u}^{(2)} - \boldsymbol{u}^{(1)}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(2)}) \cdot \boldsymbol{p}^{(2)} && \le 0\\
(1,3,1):&\quad (\boldsymbol{u}^{(3)} - \boldsymbol{u}^{(1)}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(3)}) \cdot \boldsymbol{p}^{(3)} && \le 0\\
(2,3,2):&\quad (\boldsymbol{u}^{(3)} - \boldsymbol{u}^{(2)}) \cdot \boldsymbol{p}^{(2)} + (\boldsymbol{u}^{(2)} - \boldsymbol{u}^{(3)}) \cdot \boldsymbol{p}^{(3)} && \le 0\\
(1,2,3,1):&\quad (\boldsymbol{u}^{(2)} - \boldsymbol{u}^{(1)}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{u}^{(3)} - \boldsymbol{u}^{(2)}) \cdot \boldsymbol{p}^{(2)} + (\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(3)}) \cdot \boldsymbol{p}^{(3)} && \le 0\\
(1,3,2,1):&\quad (\boldsymbol{u}^{(3)} - \boldsymbol{u}^{(1)}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{u}^{(2)} - \boldsymbol{u}^{(3)}) \cdot \boldsymbol{p}^{(3)} + (\boldsymbol{u}^{(1)} - \boldsymbol{u}^{(2)}) \cdot \boldsymbol{p}^{(2)} && \le 0
\end{aligned}$$

Specifically in this example, the criterion function is:

$$\begin{aligned}
Q(\boldsymbol{\beta}) =& \big[(\boldsymbol{X}^{(2)}\boldsymbol{\beta} - \boldsymbol{X}^{(1)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{X}^{(1)}\boldsymbol{\beta} - \boldsymbol{X}^{(2)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(2)}\big]_+^2 +\\
& \big[(\boldsymbol{X}^{(2)}\boldsymbol{\beta} - \boldsymbol{X}^{(1)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{X}^{(1)}\boldsymbol{\beta} - \boldsymbol{X}^{(3)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(3)}\big]_+^2 + \ldots\\
& \big[(\boldsymbol{X}^{(3)}\boldsymbol{\beta} - \boldsymbol{X}^{(1)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(1)} + (\boldsymbol{X}^{(2)}\boldsymbol{\beta} - \boldsymbol{X}^{(3)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(3)} + (\boldsymbol{X}^{(1)}\boldsymbol{\beta} - \boldsymbol{X}^{(2)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(2)}\big]_+^2
\end{aligned}$$

[12] All permutations such that no two permutation can be rotated into one another, where rotation means changing the start of the cycle, but the relative positions of the numbers are unchanged. For instance, the cycle (1,2,3,1) is a rotation of the cycle (2,3,1,2).

		Market $m = 1$	Market $m = 2$	Market $m = 3$
Covariates of product 1,	$\boldsymbol{X}_1^m$	$(0, -1)$	$(0, 1)$	$(1, 0)$
Covariates of product 2,	$\boldsymbol{X}_2^m$	$(0, 0)$	$(0, 0)$	$(0, 0)$
Market share of product 1,	p_1^m	0.5	0.75	0.25
Market share of product 2,	p_2^m	0.5	0.25	0.75

Table 1: Example dataset with two choices and three markets.

We provide a computational example here. Suppose we are given the dataset in Table 1. In this dataset, there are two goods available in each of the three market. Suppose we want to see if the parameters $\beta_1 = 1$, $\beta_2 = 0$ can be rejected by the dataset. At this $\boldsymbol{\beta}$, the implied utility (i.e. $u_j^m = \beta_1 X_{j1}^m + \beta_2 X_{j2}^m$) for choosing product $j = 1$ in market m is zero for markets $1, 2$; and one for market 3. We normalize the utilities from choosing product $j = 2$ to be zero in all markets. Then the criterion function $Q(\beta_1 = 1,\ \beta_2 = 0)$ is calculated as in Table 2. Intuitively, cyclical monotonicity is violated because the market share of product 1 is the lowest in market 3, yet at these parameters, the utility from choosing product 1 in market 3 is the highest.

Cycles	Violation of cyclical monotonicity
(1,2,1)	0
(1,3,1)	$(\max\{0, 0.5 - 0.25\})^2$
(2,3,2)	$(\max\{0, 0.75 - 0.25\})^2$
(1,2,3,1)	$(\max\{0, 0.75 - 0.25\})^2$
(1,3,2,1)	$(\max\{0, 0.5 - 0.25\})^2$
Sum	0.625

Table 2: $Q(\beta_1 = 1, \beta_2 = 0) = 0.625 > 0$. At these parameters, cyclical monotonicity inequalities are violated for all cycles except for the first one, the value $Q(\boldsymbol{\beta}) = 0.625$ measures the degree of violation from cyclical monotonicity at $\beta_1 = 1, \beta_2 = 0$.

2.3. Remarks

Unlike the Logit model, our modeling framework does not require the population choice probabilities to be strictly positive.[13] That is, our model allows for vectors of choice probabilities where one or many choices have zero probability of being chosen.[14] This is empirically relevant in the setting of large choice set because we often observe zero choice

[13] See Gandhi, Lu & Shi (2013)

[14] Specifically, we do not impose that ϵ has an unbounded support. Hence it is entirely possible that the subgradient of $\mathcal{G}(\boldsymbol{u})$ consists of a probability vector $\boldsymbol{p}(\boldsymbol{u})$ such that some components of $\boldsymbol{p}(\boldsymbol{u})$ are zero. As such, cyclical monotonicity inequalities are valid when some of the observed choice probabilities are zero.

probabilities, perhaps due to the way agents form their consideration set.

Moreover, the cyclical monotonicity inequalities involve differences in utilities across markets. Suppose that the vector of mean utilities for each market m is given by $\boldsymbol{U}^{(m)} = \boldsymbol{X}^{(m)}\boldsymbol{\beta} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a vector of product-specific unobservables. If ξ does not vary across markets, then when we compute the cyclical monotonicity inequalities, ξ will be differenced out. Therefore, our approach allows for certain types of unobservables. On the flipside, any market-invariant variables are not identified.[15]

The exogeneity assumption allows us to form pairwise and cyclical comparisons across markets. In some contexts, this may be reasonable. For instance, when we observe relatively high-frequency data (each market is defined as a few days or a week), then firms' pricing decision (the covariate, $\boldsymbol{X}$) may not change so frequently as to take advantage of weekly demand shocks (the error, $\boldsymbol{\epsilon}$). Price changes at a high-frequency level are likely due to predetermined marketing campaigns.

When the dimension of the choice sets increases, minimizing the criterion function in Equation 5 becomes increasingly difficult. We now describe how random projection can reduce the dimensionality of our problem.

3. Random Projection

Our approach consists of two-steps: in the first data-preprocessing step, the data matrix $\mathcal{D}$ is embedded into a lower-dimensional Euclidean space. This dimensionality reduction is achieved by premultiplying $\mathcal{D}$ with a *random projection matrix*, resulting in a projected-down data matrix $\tilde{\mathcal{D}}$ with a fewer number of rows, but the same number of columns (that is, the number of markets and covariates is not reduced, but the dimensionality of choice sets is reduced). In the second step, the estimator outlined in Equation (5) is computed using only the projected-down data $\tilde{\mathcal{D}}$.

A random projection matrix R, is a k-by-d matrix (with $k \ll d$) such that each entry $R_{i,j}$ is distributed i.i.d with mean zero and variance $\frac{1}{k}$. An attractive class of random projections is the *sparse random projection matrices* (Li et al. (2006)), in which many elements are zero with high probability.

Definition 3 (Sparse Random Projection Matrix)**:** A sparse random projection matrix is a k-by-d matrix R such that each i, j-th entry is independently and identically distributed

[15] In the BLP (Berry et al. (1995)) model, it is also possible recover the product-specific unobservables. Compared to the BLP, our framework currently does not allow for endogenous regressors and random coefficients.

according to the following discrete distribution:

$$R_{i,j} = \sqrt{\frac{s}{k}} \begin{cases} +1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ -1 & \text{with probability } \frac{1}{2s} \end{cases} \quad (s > 1).$$

By choosing a higher s, we produce sparser random projection matrices. Sparse random projection matrices are significantly faster to generate and manipulate, as well as requiring less memory to store. Gaussian random projection is also a popular class of random projections, but uniform numbers are much easier to generate than Gaussian random numbers

3.1. How random projection works

Before proceeding, we illustrate the intuitions behind random projections. We begin with a high-dimensional vector:

$$\boldsymbol{x} = (x_1, x_2, \ldots, \ldots, \ldots, x_d).$$

Then we premultiply $\boldsymbol{x}$ by the random $k \times d$ matrix $\boldsymbol{R}$ described in Definition 3, as illustrated in Figure 1. This leads to a much shorter k-dimensional vector $\tilde{\boldsymbol{x}} = \boldsymbol{R} \cdot \boldsymbol{x} = \left[\boldsymbol{R_1} \cdot \boldsymbol{x}, \boldsymbol{R_2} \cdot \boldsymbol{x}, \ldots, \boldsymbol{R_k} \cdot \boldsymbol{x}\right]'$, where $\boldsymbol{R}_i$ denotes the i-th row of $\boldsymbol{R}$. That is, each element of the lower-dimensional vector $\tilde{\boldsymbol{x}}$ is a random linear combination of elements in the high-dimensional vector $\boldsymbol{x}$. Now consider the squared Euclidean length of $\tilde{\boldsymbol{x}}$, which is random and is given as follows.

$$\|\tilde{\boldsymbol{x}}\|^2 = \underbrace{(\sum_{i=1}^{d} R_{1i} x_i)^2}_{\frac{1}{k}\|x\|^2 \text{ on average}} + \cdots + (\sum_{i=1}^{d} R_{ki} s_i)^2 \tag{6}$$

It turns out that every term in the sum on the right-hand side of Equation 6 is, on average, equal to $\frac{1}{k}\|\mathbf{x}\|^2$.[16] That is, $\|\tilde{\boldsymbol{x}}\|^2$ is a sum of k i.i.d terms, each of which has mean equal to $\frac{1}{k}\|\boldsymbol{x}\|^2$. This implies, then, that the high-dimensional vector $\boldsymbol{x}$ and the lower-dimensional projected-down vector $\widetilde{\boldsymbol{x}}$ have the *same length* on average:

$$\mathbb{E}[\|\widetilde{\boldsymbol{x}}\|^2] = \|\boldsymbol{x}\|^2$$

[16] To see this: $\mathbb{E}[(\sum_i R_{1i} x_i)^2] = \mathbb{E}[\cdots + R_{1i}^2 x_i^2 + R_{1i} x_i \, R_{1j} x_j + \ldots]$ where all the cross-terms are zero in expectation due to independence. Then since $\mathbb{E}[R_{1i}^2] x_i^2 = \frac{1}{k} x_i^2$, we have $\mathbb{E}[(\sum_i R_{1i} x_i)^2] = \frac{1}{k}(x_1^2 + \cdots + x_d^2) = \frac{1}{k}\|\boldsymbol{x}\|^2$.

- Full dimension $d = 5000$. Reduced dimension $k = 50$. Sparseness $s = 100$.
- $R_{i,j} \in \sqrt{2} \cdot \{+1, 0, -1\}$ i.i.d. with prob. $\{0.005, 0.99, 0.005\}$

$$\boldsymbol{R} := \left.\overbrace{\begin{pmatrix} 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & -1 & 0 & 0 & 0 & \dots & 0 \\ \vdots & & & & & & & & \\ 0 & 1 & \dots & 0 & 0 & 0 & -1 & \dots & 0 \end{pmatrix}}^{d = 5000 \text{ columns}} * \sqrt{2} \;\right\} k = 50 \text{ rows}$$

- Random projection: $\underbrace{\tilde{\boldsymbol{x}}}_{k \times 1} = \underbrace{\boldsymbol{R}}_{k \times d} \cdot \underbrace{\boldsymbol{x}}_{d \times 1}$

Figure 1: Example of random projection

Essentially, the defining feature of random projections is that it preserves the length of the high-dimensional vector after dimensionality-reduction.[17]

Moreover Li et al. (2006) show that:

$$\mathrm{Var}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = \frac{1}{k}\Big(2\|\boldsymbol{u} - \boldsymbol{v}\|^4 + (s-3)\sum_{j=1}^{d}(u_j - v_j)^4\Big) \tag{7}$$

When d is large, which is exactly the setting where random projection is needed, the first term in the variance formula dominates the second term. Therefore, we can set large values of s to achieve very sparse random projection, with negligible loss in efficiency. More concretely, we can set s to be as large as $\sqrt{d}$. We will see in the simulation example that when $d = 5000$, setting $s = \sqrt{d}$ implies that the random projection matrix is zero with probability 0.986 – that is, only 1.4% of the data are sampled on average. Yet we find that sparse random projection performs just as well as a dense random projection.[18] The conventional (dense) Gaussian random projection has the same variance as when $s = 3$ in the sparse random projections.

There are new developments in random projection that we could exploit to further reduce the costs of computation and data storage. For instance, Li, Mitzenmacher & Shrivastava (2014) show that we can even code the projected data using a one-bit coding scheme (i.e.

[17] For a detailed discussion, see Chapter 1 in Vempala (2000).

[18] More precisely, as shown by Li et al. (2006), is that if all fourth moments of the data to be projected-down are finite, i.e. $\mathbb{E}[u_j^4] < \infty$, $\mathbb{E}[v_j^4] < \infty$, $\mathbb{E}[u_j^2 v_j^2] < \infty$, for all $j = 1, \dots, d$, then the term $\|\boldsymbol{u} - \boldsymbol{v}\|^4$ in the variance formula (Eq. 7) dominates the second term $(s-3)\sum_{j=1}^{d}(u_j - v_j)^4$ for large d (which is precisely the setting we wish to use random projection).

just recording the signs of each number), and in some cases, there is not much accuracy loss when performing certain tasks such as similarity estimation.

3.2. Random Projection Estimator

We introduce the random projection estimator. Given the dataset $\mathcal{D} = \{(\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(M)}, \boldsymbol{p}^{(M)})\}$, define the *projected-down* dataset by $\tilde{\mathcal{D}}_k = \{(\tilde{\boldsymbol{X}}^{(1)}, \tilde{\boldsymbol{p}}^{(1)}), \ldots, (\tilde{\boldsymbol{X}}^{(M)}, \tilde{\boldsymbol{p}}^{(M)})\}$, where $(\tilde{\boldsymbol{X}}^{(m)}, \tilde{\boldsymbol{p}}^{(m)}) = (R\boldsymbol{X}^{(m)}, R\boldsymbol{p}^{(m)})$ for all markets m, and R being a sparse $k \times d$ random projection matrix as in Definition 3. Note that the same random projection matrix is used to project down all markets.

Definition 4 (Random projection estimator)**:** The random projection estimator is defined as $\tilde{\boldsymbol{\beta}}_k \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\|=1} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$, where $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ is the criterion function in Equation (5) in which the input data is $\tilde{\mathcal{D}}_k$. ■

The projected-down dataset $\tilde{\mathcal{D}}_k$ has k number of rows, where the original dataset has a larger number of rows, d. Note that the identities of the markets and covariates (i.e. the columns of the data matrix) are unchanged in the reduced-dimension data matrix; as a result, the same projected-down dataset can be used to estimate different utility/model specifications with varying combination of covariates and markets.

We will benchmark the random projection estimator with the estimator $\hat{\boldsymbol{\beta}} \in \operatorname{argmin} Q(\boldsymbol{\beta}, \mathcal{D})$, where $Q(\boldsymbol{\beta}, \mathcal{D})$ is the criterion function in Equation (5) in which the original data $\mathcal{D}$ is used as input. In the next section, we will prove convergence of the random projection estimator to the benchmark estimator using the original data, as k grows large. Here we provide some intuition and state some preliminary results for this convergence result.

Recall from the previous section that the Euclidean distance between two vectors are preserved in expectation as these vectors are projected-down into a lower-dimensional Euclidean space. In order to exploit this feature of random projection for our estimator, we rewrite the estimating inequalities – based on cyclical monotonicity – in terms of Euclidean norms, using Definition 2.

From Definition 2, we can rewrite the estimator in (5) as $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$ where the criterion function is defined as the sum of squared violations of the cyclical monotonicity inequalities:

$$Q(\boldsymbol{\beta}) = \sum_{\text{all cycles in data } \mathcal{D}; L \geq 2} \left[\sum_{m=2}^{L+1} \left(\|\boldsymbol{X}^{(a_l)}\boldsymbol{\beta} - \boldsymbol{p}^{(a_l)}\|^2 - \|\boldsymbol{X}^{(a_l)}\boldsymbol{\beta} - \boldsymbol{p}^{(a_{l-1})}\|^2 \right) \right]_+^2 \qquad (8)$$

To see another intuition behind the random projection estimator, we introduce the Johnson-Lindenstrauss Lemma. This lemma states that there exists a linear map (which can be found by drawing different random projection matrices) such that there is a low-distortion embedding. There are different versions of this theorem; we state a typical one:

Lemma 1 (Johnson-Lindenstrauss)**.** *Let* $\delta \in (0, \frac{1}{2})$*. Let* $\mathcal{U} \subset \mathbb{R}^d$ *be a set of* C *points, and* $k = O(\log C/\delta^2)$*. There exists a linear map* $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^k$ *such that for all* $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$*:*

$$(1-\delta)\|\boldsymbol{u}-\boldsymbol{v}\|^2 \leq \|\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{v})\|^2 \leq (1+\delta)\|\boldsymbol{u}-\boldsymbol{v}\|^2.$$

The crucial insight of this lemma lies in the proof.[19] The proof is probabilistic, and demonstrates that if we take $\boldsymbol{f}$ to be a random projection matrix, then there is a non-zero probability such that $\|\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{v})\|^2$ satisfies the error bounds stated in the Lemma. Hence, there exists a realization of $\boldsymbol{f}$ that would serve as the appropriate linear mapping. For this reason, the Johnson-Lindenstrauss Lemma has become a term that collectively represents random projection methods, even when the lemma itself is not used.

As the statement of the Lemma makes clear, the reduced-dimension k controls the trade-off between tractability and precision. Moreover, implicit in the Lemma is that we can find an embedding of points from d-dimension to k-dimension, independent of how large d is.[20]

The feature that the cyclical monotonicity inequalities can be written in terms of Euclidean norms between vectors justifies the application of random projection to our estimator. In contrast, the "rank-order" inequalities, which underlie the maximum score approach to semiparametric multinomial choice estimation,[21] cannot be rewritten in terms in terms of Euclidean norms between data vectors, and hence random projection cannot be used for those inequalities.

3.3. Convergence

In this section we show that, for any given data $\mathcal{D}$, the random projection estimator computed using the projected-down data $\tilde{\mathcal{D}}_k = R \cdot \mathcal{D}$ converges in probability to the corresponding estimator computed using the original data $\mathcal{D}$, as k grows large.

[19] Proofs of the Johnson-Lindenstrauss Lemma can be found in, among others, Achlioptas (2003); Dasgupta & Gupta (2003); Vempala (2000).

[20] Moreover according to Li et al. (2006),"the Johnson-Lindenstrauss lemma is conservative in many applications because it was derived based on Bonferroni correction for multiple comparisons." That is, the magnitude for k in the statement of the Lemma is a worst-case scenario, and larger than necessary in many applications. This is seen in our computational simulations below, where we find that small values for k still produce good results.

[21] For instance, Manski (1985), Fox (2007). The rank-order property makes pairwise comparisons of choices *within* a given choice set, and state that, for all $i, j \in \mathcal{C}$, $p_i(\boldsymbol{u}) > p_j(\boldsymbol{u})$ iff $u_i > u_j$.

The main insight here is that the error introduced by random projections disappears as k (the reduced dimension) increases. Our random projection estimator will obtain the same values as (i.e. be consistent with) the values that would be produced using the original, high-dimensional dataset, as k increases. We will further see in simulations that even with a modest k, our random projection estimates are close to the original estimates.

In order to highlight the random projection aspect of our estimator, we assume that the market shares and other data variables are observed without error.[22] Hence, given the original data $\mathcal{D}$, the criterion function $Q(\boldsymbol{\beta}, \mathcal{D})$ is deterministic, while the criterion function $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ is random solely due to the random projection procedure.

All proofs are provided in Appendix A.2. We first show that the random-projected criterion function converges uniformly to the unprojected criterion function in Theorem 1.

Theorem 1 (Uniform convergence of criterion function)**.** *For any given $\mathcal{D}$, we have $\sup_{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\|=1} |Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D})| \xrightarrow{p} 0$, as k grows.*

Essentially, from the defining features of the random projection matrix R, we can argue that $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in probability to $Q(\boldsymbol{\beta}, \mathcal{D})$, *pointwise* in $\boldsymbol{\beta}$. Then, because $Q(\boldsymbol{\beta}, \mathcal{D})$ is convex in $\boldsymbol{\beta}$ (which we will also show), we can invoke the Convexity Lemma from Pollard (1991), which says that pointwise and uniform convergence are equivalent for convex random functions.

Finally, under the assumption that the deterministic criterion function $Q(\boldsymbol{\beta}, \mathcal{D})$ (i.e. computed without random projection) admits an identified set, then the random projection estimator converges in a set-wise sense to the same identified set. Convergence of the set estimator here means convergence in the half-Hausdorff distance, where the half-Hausdorff distance between two sets is defined as $d(X, Y) = \sup_{x \in X} \inf_{y \in Y} \|x - y\|$, and $\|\cdot\|$ is the Euclidean norm.

Assumption 1 (Existence of an identified set Θ^*)**.** Denote Θ as the domain of the parameters, i.e. $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| = 1\}$. There exists a set $\Theta^* \subset \Theta$ such that $\sup_{\boldsymbol{\beta} \in \Theta^*} Q(\boldsymbol{\beta}, \mathcal{D}) = \inf_{\boldsymbol{\beta} \in \Theta} Q(\boldsymbol{\beta}, \mathcal{D})$, and $\forall \nu > 0$, $\inf_{\boldsymbol{\beta} \notin B(\Theta^*, \nu)} Q(\boldsymbol{\beta}, \mathcal{D}) > \sup_{\boldsymbol{\beta} \in \Theta^*} Q(\boldsymbol{\beta}, \mathcal{D})$, where $B(\Theta^*, \nu)$ denotes a union of open balls of radius ν each centered on each element of Θ^*.

Theorem 2. *Suppose that Assumption 1 holds. For any given $\mathcal{D}$, the random projection estimator $\tilde{\Theta}_k = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^b : ||\boldsymbol{\beta}||=1} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in half-Hausdorff distance to the*

[22] This is not an unreasonable starting point, because in most typical applications of aggregate discrete-choice demand models, the observed market shares are purchase frequencies computed from very large sample size, and hence sampling error should be rather negligible.

identified set Θ^* *as* k *grows, i.e.* $\sup_{\boldsymbol{\beta}\in\tilde{\Theta}_k}\inf_{\boldsymbol{\beta}'\in\Theta^*}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|\xrightarrow{p}0$ *as* k *grows.*

Theorem 2 above shows that the approximation error due to random projection disappears as k increases, assuming that d is fixed. This result illustrates that our estimator will converge to the values of the unprojected estimator as k increases.

We can achieve a stronger result where d is **not** fixed, but is allowed to grow with k. In the Appendix A.3, we show using Lemma 2 from Li et al. (2006) that the bound $\mathrm{Var}(\|R\boldsymbol{u}-R\boldsymbol{v}\|^2)=O(\frac{1}{k})$ does not depend on d. Therefore we can let both d and k grow to infinity, the convergence result of Theorem 2 would still hold. There is no restriction on what the original dimension is, d can be very large or moderate in relation to k. Hence our estimator is suited for applications with very large choice set, as the original dimension d does not affect the performance of the estimator. As long as the reduced dimension k is not too small, the random projection estimator would approximate well.

4. Simulation Examples

In this section, we show simulation evidence that random projection performs well in practice. In these simulations, the sole source of randomness is the random projection matrices (in Section 4.3, we relax this assumption). This allows us to starkly examine the noise introduced by random projections, and how the performance of random projections varies as we change k, the reduced dimensionality. Our result shows that the error introduced by random projection is small, even when the reduced dimension k is small.

4.1. Setup

We consider projecting down from d to k. Recall that d is the number of choices in our context. There are $n = 30$ markets. The utility that an agent in market m receives from choice j is $U_j^{(m)}=\beta_1 X_{1,j}^{(m)}+\beta_2 X_{2,j}^{(m)}$, where $X_{1,j}^{(m)}\sim N(1,1)$ and $X_{2,j}^{(m)}\sim N(-1,1)$ independently across all choices j and markets m.[23]

We normalize the parameters $\beta=(\beta_1,\beta_2)$ such that $\|\beta\|=1$. This is achieved by parameterizing β using polar coordinates: $\beta_1=\cos\theta$ and $\beta_2=\sin\theta$, where $\theta\in[0,2\pi]$. The true parameter is $\theta_0=0.75\pi=2.3562$.

To highlight a distinct advantage of our approach, we choose a distribution of the error term that is neither exchangeable nor belongs to the generalized extreme value family.

[23] We also considered two other sampling assumptions on the regressors, and found that the results are robust to: (i) strong brand effects: $X_{l,j}^{(m)}=X_{l,j}+\eta_{l,j}^{(m)}$, $l=1,2$, where $X_{1,j}\sim N(1,0.5)$, $X_{2,j}\sim N(-1,0.5)$, and $\eta_{l,j}^{(m)}\sim N(0,1)$; (ii) strong market effects: $X_{l,j}^{(m)}=X_l^{(m)}+\eta_{l,j}^{(m)}$, $l=1,2$, where $X_1^{(m)}\sim N(1,0.5)$, $X_2^{(m)}\sim N(-1,0.5)$, and $\eta_{l,j}^{(m)}\sim N(0,1)$.

Specifically, we let the additive error term be a MA(2) distribution where errors are serial correlated in errors across products. To summarize, the utility that agent in market m derives from choice j is $U_j^{(m)} + \epsilon_j^{(m)}$, where $\epsilon_j^{(m)} = \frac{1}{3}\sum_{l=0}^{3} \eta_{j+l}^{(m)}$, and $\eta_j^{(m)}$ is distributed i.i.d with $N(0,1)$.

Using the above specification, we generate the data $\mathcal{D} = \{(\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(M)}, \boldsymbol{p}^{(M)})\}$ for $n = 30$ markets, where $\boldsymbol{p}^{(m)}$ corresponds to the d-dimensional vector of simulated choice probabilities for market m: the j-th row of $\boldsymbol{p}^{(m)}$ is $p_j^{(m)} = \Pr\left(U_j^{(m)} + \epsilon_j^{(m)} > U_{-j}^{(m)} + \epsilon_{-j}^{(m)}\right)$. We then perform random projection on $\mathcal{D}$ to obtain the projected-down dataset $\tilde{\mathcal{D}} = \{(\tilde{\boldsymbol{X}}^{(1)}, \tilde{\boldsymbol{p}}^{(1)}), \ldots, (\tilde{\boldsymbol{X}}^{(M)}, \tilde{\boldsymbol{p}}^{(M)})\}$. The random projection matrix in Definition 3 is parameterized by s. We set $s = \sqrt{d}$, which corresponds to very sparse random projection matrices. We restrict to cycles of length 2 and 3 in computing Equation 5; however, we find that even using cycles of length 2 did not noticeably change the result.

Table 3 shows the main result for this section. In the table, the rows correspond to different designs where the dimension of the dataset is projected down from d to k. For each design, we estimate the model using 100 independent realizations of the random projection matrix. We report the means of the upper and lower bounds of the estimates, as well as their standard deviations. We also report the interval spans by the 25th percentile of the lower bounds as well as the 75th percentile of the upper bounds.

The results indicate that, in most cases, optimization of the randomly-projected criterion function $Q(\boldsymbol{\beta}, \mathcal{D}_k)$ yields a unique minimum. For instance, in the fourth row of Table 9 (when compressing from $d = 5000$ to $k = 100$), we see that the projected criterion function is always uniquely minimized[24] (across all 100 replications). Moreover the average point estimate for θ is equal to 2.3878, where the true value is 2.3562.

Design	mean LB (s.d.)	mean UB (s.d.)	25th LB, 75th UB
$d = 100, k = 10$	2.2772 (0.2461)	2.2772 (0.2461)	[2.0813, 2.4740]
$d = 500, k = 100$	2.2970 (0.2634)	2.3972 (0.2697)	[2.1049, 2.6075]
$d = 1000, k = 100$	2.3014 (0.2277)	2.3051 (0.2262)	[2.1284, 2.4504]
$d = 5000, k = 100$	2.3878 (0.2891)	2.3878 (0.2891)	[2.2305, 2.5918]
$d = 5000, k = 500$	2.2982 (0.3035)	2.5017 (0.2950)	[2.0813, 2.7332]

Table 3: Random projection estimator with sparse random projections, $s = \sqrt{d}$. Replicated 100 times using independently realized sparse random projection matrices (where $s = \sqrt{d}$ in Definition 3). The true value of θ is 2.3562.

[24] The fact that the criterion function is uniquely minimized is not surprising, and occurs often in the moment inequality literature; the random projection procedure introduces noise into the projected inequalities so that, apparently, there are no values of the parameters $\boldsymbol{\beta}$ which jointly satisfy all the projected inequalities, leading to a unique minimizer for the projected criterion function.

Additionally, we conduct the same simulation exercise, except that we use dense random projection matrices instead. In Table 9 in the appendix, we set $s = 1$. We see that there is no noticeable difference in performance, but sparse random projections are more computationally efficient

4.2. Computational details

Generating and manipulating a very large, but sparse random projection matrix can be implemented efficiently in many standard statistical programs. For instance in MATLAB, we can generate sparse matrices using the command "sprand", which are stored and treated as a different object than the usual matrices. Besides requiring less memory to store, multiplying a sparse matrix with a dense matrix is also computationally efficient.

In Table 4, we show that the entire process of generating a random projection matrix, *and* using it to project down a large data matrix is trivially fast (0.038 seconds), even for dimensions of choice set up to 100,000.[25] We set the sparseness parameter $s = \sqrt{d}$ as in Definition 3.

d	Time (seconds)
20,000	0.0109
40,000	0.0182
60,000	0.0316
80,000	0.0362
100,000	0.0384

Table 4: Time for the entire process of generating a sparse random projection matrix of size $k \times d$, **and** using it to project down a large data matrix consisting of d rows and 90 columns.[26] We set $k = 100$.

Because performing random projection is very fast, we further see in Table 5 that our estimator takes a virtually constant amount of time to run as the dimension of choice set increases from 2000 to 7000. In contrast, without random projection, estimation takes an increasing amount of time as d increases.

The computational saving is significant. To put in perspective, say we want to subsample (or bootstrap), and minimize the criterion function 100 times as in Chernozhukov, Hong & Tamer (2007) to obtain the confidence region (see Section 4.3 below). At $d = 7000$, this

[25] All computation here is ran on MATLAB using a 2.0 GHz quad-core Intel Core i7 processor (late 2013 MacBook Pro). Time is calculated as median elasped CPU time over 10 runs.

[26] Each entry of the data matrix is drawn i.i.d. from Uniform$[0, 1]$. 90 columns would correspond to 30 markets, each market consists of a vector of choice probability and two vectors of covariates.

would take about 23 hours without random projection, bearing in mind there are only two parameters to estimate. With random projection, this would take about 1 hour.

d	Time (seconds)	
	(2) random projection	(3) without random projection
2,000	36.1	128.6
3,000	36.4	170.6
4,000	36.5	593.7
5,000	36.6	662.6
6,000	36.2	756.9
7,000	35.9	826.0

Table 5: In column (2), time takes to run the entire process: generate random projection and compute the minima of the criterion function using the projected-down data. In column (3), time takes to compute the minima of the criterion function using the original data. The setup here is the same as the simulation setup (Section 4.1), where the criterion function only has one variable to minimize over. As before, the reduced dimension is $k = 100$.

4.3. Estimating the confidence region

We briefly describe the Chernozhukov et al. (2007) (CHT) procedure for estimating the confidence region.[27] Intuitively, this procedure bootstraps the data to obtain a critical value c^*. Then, the confidence region is those parameters that lie within the c^*-level set of the criterion function, i.e. $\{\boldsymbol{\beta} : Q(\boldsymbol{\beta}; \mathcal{D}) < c^*\}$.

First, let $c_0 = \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}; \mathcal{D})$ and $\mathcal{C}_0 = \{\boldsymbol{\beta} : Q(\boldsymbol{\beta}; \mathcal{D}) \leq c_0\}$, which corresponds to the initial estimates. Second, draw 100 bootstrapped dataset (described in the next paragraph), which we denote by $\mathcal{D}^{(r)}$, $r = 1, \ldots, 100$. Third, let c_1 be the 95th percentile of $\{\max_{\boldsymbol{\beta} \in \mathcal{C}_0} Q(\boldsymbol{\beta}; \mathcal{D}^{(r)}) - \min_{\tilde{\boldsymbol{\beta}}} Q(\tilde{\boldsymbol{\beta}}; \mathcal{D}^{(r)})\}_{r=1,\ldots,100}$. Finally, after obtaining c_1, the 95% confidence region is given by $\mathcal{C}_1 = \{\boldsymbol{\beta} : Q(\boldsymbol{\beta}; \mathcal{D}) - \min_{\tilde{\boldsymbol{\beta}}} Q(\tilde{\boldsymbol{\beta}}; \mathcal{D}) \leq c_1\}$.[28] As shown in the previous section, this procedure is computationally intensive without random projection. Hence to combine the CHT procedure with random projections, we draw a random projection matrix R, and use it to project-down all the bootstrapped dataset. That is, we replace $\mathcal{D}$ and $\mathcal{D}^{(r)}$ above with $\tilde{\mathcal{D}} = R\mathcal{D}$ and $\tilde{\mathcal{D}}^{(r)} = R\mathcal{D}^{(r)}$ for all r.

There are many ways to construct the bootstrapped dataset, the idea is for each bootstrapped dataset to lie within the sampling variation of the original dataset. One common

[27] This procedure is also used in Ciliberto & Tamer (2009).

[28] Given the assumptions in Chernozhukov et al. (2007), the true identified set θ_I that minimizes the population criterion function would satisfy $\lim_{n\to\infty} \Pr(\theta_I \in \mathcal{C}_1) \geq 0.95$.

method is to assume that we have access to the individual-level observations such that the aggregate market shares or choice probabilities are estimated from these observations. Hence, each bootstrapped dataset is constructed by sampling the individuals with replacement and re-calculate the choice probabilities.

Now we relax the earlier assumption that there is no sampling error in our dataset (otherwise the setup is identical to the one before in Section 4.1). In particular we now allow for sampling error in the market share or choice probabilities. Thus far, we have constructed the choice probabilities by aggregating the choices of I consumers: $p_j^{(m)} = \sum_{i=1}^{I} \frac{1}{I} \mathbb{1}(U_j^{(m)} + \epsilon_{j,i}^{(m)} > U_{-j}^{(m)} + \epsilon_{-j,i}^{(m)})$ for a large number I. Here, we set I to be a smaller number, i.e. $I = d$. We then generate 100 independent samples of the choice probabilities, and use each of these to construct the bootstrapped sample. Finally, we compute the CHT confidence region.

The result is reported in Table 6. For each design, we report the random projection estimate. Unlike Table 3 before, we computed the random projection estimator once per design, using a single draw of the random projection. Now the last column reports the CHT confidence regions associated with these estimates. These are the 95% confidence regions for the random projection estimates when we allow for sampling error in the choice probabilities.

From Table 6, we see that the 95% confidence region for the random projection estimates is very narrow. This suggests that sampling error in the market shares does not severely affect our random projection estimator (at least in the context of our simulation example).

Design	RP estimates	CHT confidence region
$d = 100, k = 10$	1.9478	$[1.8535, 2.0263]$
$d = 500, k = 100$	2.3248	$[2.2462, 2.3876]$
$d = 1000, k = 100$	2.1834	$[2.1363, 2.2462]$
$d = 5000, k = 100$	2.0933	$[2.0681, 2.1121]$
$d = 5000, k = 500$	2.3383	$[2.2818, 2.3886]$

Table 6: The second column shows the random projection estimates for various designs (each estimate is computed using a single draw of random projection). The last column shows the CHT confidence region associated with this estimate. The actual value of the parameter is 2.3562.

5. Empirical Application: an aggregate demand model incorporating both store and brand choices

For our empirical application, we use supermarket scanner data made available by the Chicago-area Dominicks supermarket chain.[29] Dominick's operated a chain of grocery stores across the Chicago area, and the database recorded sales information on many product categories, at the store and week level. For this application, we look at the soft drinks category.

For our choice model, we consider a model in which consumers choose both the type of soft drink, as well as the *store* at which they make their purchase. Such a model of joint store and brand choice allows consumers not only to change their brand choices, but also their store choices, in response to across-time variation in economic conditions. For instance, Coibion, Gorodnichenko & Hong (2015) is an analysis of supermarket scanner data which suggests the importance of "store-switching" in dampening the effects of inflation in posted store prices during recessions.

Such a model of store and brand choice also highlights a key benefit of our semiparametric approach. A typical parametric model which would be used to model store and brand choice would be a nested logit model, in which the available brands and stores would belong to different tiers of nesting structure. However, one issue with the nested logit approach is that the results may not be robust to different assumptions on the nesting structure– for instance, one researcher may nest brands below stores, while another researcher may be inclined to nest stores below brands. These two alternative specifications would differ in how the joint distribution of the utility shocks between brands at different stores are modeled, leading to different parameter estimates. Typically, there are no *a priori* guides on the correct nesting structure to impose.[30]

In this context, a benefit of our semiparametric approach is that we are *agnostic* as to the joint distribution of utility shocks; hence our approach accommodates both models in which stores are in the upper nest and brands in the lower nest, or vice versa, or any other model in which the stores or brands could be divided into further sub-nests.

We have $n = 15$ markets, where each market corresponds to a distinct two-weeks interval between October 3 1996 to April 30 1997, which is the last recorded date. We include sales at eleven Dominicks supermarkets in north-central Chicago, as illustrated in Figure 2. Among these eleven supermarkets, most are classified as premium-tier stores, while two are medium-tier stores (distinguished by dark black spots in Figure 2); stores in different

[29] This dataset has previously been used in many papers in both marketing and economics; see a partial list at http://research.chicagobooth.edu/kilts/marketing-databases/dominicks/papers.

[30] Because of this, Hausman & McFadden (1984) have developed formal econometric specification tests for the nested logit model.

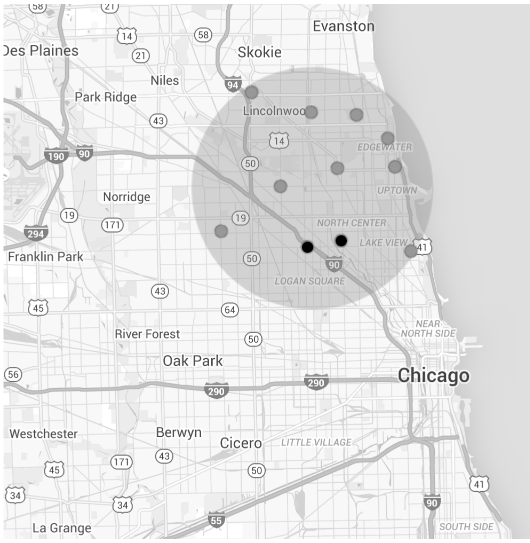

Figure 2: Location of the 11 Dominick's branches as indicated by spots.

Radius of the circle is 4 miles. The darker spots are Dominick's medium-tier stores, the rest are high-tiers.

tiers sell different ranges of products.

Variables	Definition	Summary statistics
price$_{ij}$	The average price of the store-upc j at market i	Mean: \$2.09, s.d: \$1.77
bonus$_{ij}$	The fraction of weeks in market i for which store-upc j was on sale as a bonus or promotional purchase; for instance "buy-one-get-one-half-off" deals	Mean: 0.27, s.d: 0.58
quantity$_{ij}$	Total units of store-upc j sold in market i	Mean: 60.82, s.d: 188.37
holiday$_{ij}$	A dummy variable indicating the period spanning 11/14/96 to 12/25/96, which includes the Thanksgiving and Christmas holidays	6 weeks (3 markets)
medium_tier$_{ij}$	Medium, non-premium stores.†	2 out of 11 stores
d	Number of store-upc	3059

Table 7: Summary statistics

Number of observations is 45885=3059 upcs × 15 markets (2-week periods). † denotes stores in the same tier, which share similar product selection, and also pricing to a certain extent.

Our store and brand choice model consists of $d = 3059$ choices, each corresponds to a unique store and universal product code (UPC) combination. We also define an outside option, for a total of $d = 3060$ choices.[31]

The summary statistics for our data sample are in Table 7.

Covariates:	Specification:			
	(A)	(B)	(C)	(D)
price	-0.6982	-0.9509	-0.7729	-0.4440
	$[-0.9420, -0.3131]$	$[-0.9869, -0.7874]$	$[-0.9429, -0.4966]$	$[-0.6821, -0.2445]$
bonus		0.0580	0.0461	0.0336
		$[-0.0116, 0.1949]$	$[0.0054, 0.1372]$	$[0.0008, 0.0733]$
price × bonus		-0.1447	-0.0904	-0.0633
		$[-0.4843, 0.1123]$	$[-0.3164, 0.0521]$	$[-0.1816, 0.0375]$
holiday	0.0901		0.0661	0.0238
	$[-0.0080, 0.2175]$		$[-0.0288, 0.1378]$	$[-0.0111, 0.0765]$
price × holiday	-0.6144		-0.3609	-0.1183
	$[-0.9013, -0.1027]$		$[-0.7048, -0.0139]$	$[-0.2368, -0.0164]$
price × medium_tier				0.4815
				$[-0.6978, 0.8067]$

Table 8: Random projection estimates, dimensionality reduction from $d = 3059$ to $k = 300$.
First row in each entry present the median coefficient, across 100 random projections. Second row presents the 25-th and 75-th percentile among the 100 random projections. We use cycles of length 2 and 3 in computing the criterion function (Equation 5).

Table 8 presents the estimation results. As in the simulation results in Section 4, we ran 100 independent random projections, and thus obtained 100 sets of parameter estimates for each model specification. The results reported in Table 8 are the medians across these runs. Since no location normalization is imposed for the error terms, we do not include constants in any of the specifications. For estimation, we used cycles of length of length 2 and 3.[32]

Across all specifications, the price coefficient is strongly negative. The *holiday* indicator has a positive (but small) coefficient, suggesting that, all else equal, the end-of-year holidays are a period of peak demand for soft drink products.[33] In addition, the interaction between *price* and *holiday* is strongly negative across specifications, indicating that households are more price-sensitive during the holiday season.

How large is this effect? Consider a soft drink product priced initially at \$1.00 with no promotion. The parameter estimates for Specification (C) suggest that during the holiday

[31] The outside option is constructed as follows: first we construct the market share p_{ij} as $p_{ij} = quantity_{ij}/custcoun_i$, where $quantity_{ij}$ is the total units of store-upc j sold in market i, and $custcoun_i$ is the total number of customers visiting the 11 stores and purchasing something at market i. The market share for market i's outside option is then $1 - \sum_{j=1}^{3093} p_{ij}$.

[32] The result did not change in any noticeable when we vary the length of the cycles used in estimation.

[33] cf. Chevalier, Kashyap & Rossi (2003). Results are similar if we define the holiday period to extend into January, or to exclude Thanksgiving.

period, households' willingness-to-pay for this product falls as much as if the price for the product increases by \$0.27 during non-holiday periods.[34]

We also obtain a positive sign on *bonus*, and the negative sign on the interaction *price* × *bonus* across all specifications, although their magnitudes are small, and there is more variability in these parameters across the different random projections. We see that discounts seem to make consumers more price sensitive (i.e. make the price coefficient more negative). Since any price discounts will be captured in the *price* variable itself, the *bonus* coefficients capture additional effects that the availability of discounts has on behavior, beyond price. Hence, the negative coefficient on the interaction *price* × *bonus* may be consistent with a bounded-rationality view of consumer behavior, whereby the availability of discount on a brand draws consumers' attention to its price, making them more aware of a product's exact price once they are aware that it is on sale.

In specification (D), we introduce the store-level covariate *medium-tier*, interacted with *price*. However, the estimates of its coefficient are noisy, and vary widely across the 100 random projections. This is not surprising, as *medium-tier* is a market-invariant variable and, apparently here, interacting it with price still does not result in enough variation for reliable estimation.

6. Conclusion

In this paper, we use random projection – a technique for solving high-dimensional problems in machine learning – for estimating multinomial-choice models where agents face a large number of choices. This arises in many empirical applications, for instance, when agents are allowed to choose combinations of many choices, and as a result, the dimension of the discrete-choice problem becomes exponentially large.

Our estimation procedure takes two steps. First, the high-dimensional choice data are projected (embedded stochastically) into a lower-dimensional Euclidean space. This procedure is justified via the Johnson-Lindenstrauss Lemma, which shows that the pairwise distances between data points are preserved after random projection. Because the random projection matrices are very sparse, this process of generating random projection matrices and projecting down large dataset is extremely fast. In the second step, estimation proceeds using the cyclical monotonicity inequalities implied by the multinomial choice model. By using these inequalities for estimation, we avoid making explicit distributional assumptions regarding the random utility errors; hence, our estimator is semi-parametric. The

[34] $-0.77\alpha = 0.0661 - (0.77 + 0.36)\alpha(1 + 0.27)$, where $\alpha = -0.1161$ is a scaling factor we used to scale the price data so that the price vector has the same length as the *bonus* vector. The rescaling of data vectors is without loss of generality, and improves the performance of random projection by Eq. (7).

estimator works well in computational simulations and in an application to a real-world supermarket scanner dataset.

We are currently considering several extensions. First, we are undertaking another empirical application in which consumers can choose among bundles of brands, which would thoroughly leverage the benefits of our random projection approach. Second, another benefit of random projection is that it preserves privacy, in that the researcher no longer needs to handle the original dataset but rather a "jumbled-up" random version of it.[35] We are currently exploring additional applications of random projection for econometric settings in which privacy may be an issue.

A. Appendix

A.1. Equivalence of alternative representation of cyclical monotonicity

Here we show the equivalence of Equations (1) and (3), as two alternative statements of the cyclical monotonicity inequalities. We begin with the second statement (3). We have

$$\sum_{l=2}^{L+1} \|\boldsymbol{u}^l - \boldsymbol{p}^l\|^2 = \sum_{l=2}^{L+1}\sum_{j=1}^{d} \left(u_j^l - p_j^l\right)^2 = \sum_{l=2}^{L+1}\left[\sum_{j=1}^{d}(u_j^l)^2 + \sum_{j=1}^{d}(p_j^l)^2 - 2\sum_{j=1}^{d} u_j^l p_j^l\right].$$

Similarly

$$\sum_{l=2}^{L+1} \|\boldsymbol{u}^l - \boldsymbol{p}^{l-1}\|^2 = \sum_{l=2}^{L+1}\sum_{j=1}^{d} \left(u_j^l - p_j^{l-1}\right)^2 = \sum_{l=2}^{L+1}\left[\sum_{j=1}^{d}(u_j^l)^2 + \sum_{j=1}^{d}(p_j^{l-1})^2 - 2\sum_{j=1}^{d} u_j^l p_j^{l-1}\right].$$

In the previous two displayed equations, the first two terms cancel out. By shifting the l indices forward we have:

$$\sum_{l=2}^{L+1}\sum_{j=1}^{d} u_j^l p_j^{l-1} = \sum_{l=1}^{L}\sum_{j=1}^{d} u_j^{l+1} p_j^l.$$

Moreover, by definition of a cycle that $u_j^{L+1} = u_j^1$, $p_j^{L+1} = p_j^1$, we then have:

$$\sum_{l=2}^{L+1}\sum_{j=1}^{d} u_j^l p_j^l = \sum_{l=1}^{L}\sum_{j=1}^{d} u_j^l p_j^l$$

Hence

$$\sum_{l=2}^{L+1}\left(\|\boldsymbol{u}^l - \boldsymbol{p}^l\|^2 - \|\boldsymbol{u}^l - \boldsymbol{p}^{l-1}\|^2\right) = 2\sum_{l=1}^{L}\sum_{j=1}^{d}\left(u_j^{l+1}p_j^l - u_j^l p_j^l\right) = 2\sum_{l=1}^{L}(\boldsymbol{u}^{l+1} - \boldsymbol{u}^l)\cdot\boldsymbol{p}^l$$

[35] cf. Heffetz & Ligett (2014).

Therefore, cyclical monotonicity of Equation (1) is satisfied if and only if this formulation of cyclical monotonicity in terms of Euclidean norms is satisfied.

□

A.2. Proof of Theorems in Section 3.3

We first introduce two auxiliary lemmas.

Lemma 2 (Convexity Lemma, Pollard (1991))**.** *Suppose $A_n(s)$ is a sequence of convex random functions defined on an open convex set S in $\mathbb{R}^d$, which converges in probability to some $A(s)$, for each $s \in S$. Then $\sup_{s\in K} |A_n(s) - A(s)|$ goes to zero in probability, for each compact subset K of S.*

Lemma 3. *The criterion function $Q(\boldsymbol{\beta}, \mathcal{D})$ is convex in $\boldsymbol{\beta} \in \mathbb{B}$ for any $\mathcal{D}$, where $\mathbb{B}$ is an open convex subset of $\mathbb{R}^b$.*

Proof. We want to show that $Q(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}') \le \lambda Q(\boldsymbol{\beta}) + (1-\lambda)Q(\boldsymbol{\beta}')$, where $\lambda \in [0,1]$, and we suppress the dependence of Q on the data $\mathcal{D}$.

$$
\begin{aligned}
&Q(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}') \\
&= \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}')\cdot \boldsymbol{p}^{(a_l)}\right]_+^2 \\
&= \sum_{\text{all cycles in data } \mathcal{D}} \left[\lambda\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}\cdot \boldsymbol{p}^{(a_l)} + (1-\lambda)\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}'\cdot \boldsymbol{p}^{(a_l)}\right]_+^2 \\
&\le \sum_{\text{all cycles in data } \mathcal{D}} \left\{\lambda\left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}\cdot \boldsymbol{p}^{(a_l)}\right]_+ + (1-\lambda)\left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}'\cdot \boldsymbol{p}^{(a_l)}\right]_+\right\}^2 \qquad (9) \\
&\le \lambda \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}\cdot \boldsymbol{p}^{(a_l)}\right]_+^2 + \\
&\quad (1-\lambda) \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)}\right)\boldsymbol{\beta}'\cdot \boldsymbol{p}^{(a_l)}\right]_+^2 \qquad (10) \\
&= \lambda Q(\boldsymbol{\beta}) + (1-\lambda)Q(\boldsymbol{\beta}')
\end{aligned}
$$

Inequality 9 above is due to the fact that $\max\{x, 0\} + \max\{y, 0\} \ge \max\{x + y, 0\}$ for all $x, y \in \mathbb{R}$. Inequality 10 holds from the convexity of the function $f(x) = x^2$. □

Theorem 1 (Uniform convergence of criterion function) For any given $\mathcal{D}$, we have $Q(\cdot, \tilde{\mathcal{D}}_k)$ converges uniformly to $Q(\cdot, \mathcal{D})$ as $k \to \infty$. That is, $\sup_{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\|=1} \left| Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D}) \right| \xrightarrow{p} 0$, as $k \to \infty$.

Proof of Theorem 1: Recall from Equation (7) that for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, and for the class of k-by-d random projection matrices, R, considered in Definition 3, we have:

$$\mathbb{E}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = \|\boldsymbol{u} - \boldsymbol{v}\|^2 \tag{11}$$

$$\text{Var}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = O\left(\frac{1}{k}\right) \tag{12}$$

Therefore by Chebyshev's inequality, $\|R\boldsymbol{u} - R\boldsymbol{v}\|^2$ converges in probability to $\|\boldsymbol{u} - \boldsymbol{v}\|^2$ as $k \to \infty$. It follows that for a given $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{p}$, we have $\|\tilde{\boldsymbol{X}}\boldsymbol{\beta} - \tilde{\boldsymbol{p}}\|^2 \to_p \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{p}\|^2$, where $\tilde{\boldsymbol{X}} = R\boldsymbol{X}$ and $\tilde{\boldsymbol{p}} = R\boldsymbol{p}$ are the projected-down versions of $\boldsymbol{X}$ and $\boldsymbol{p}$. Applying the Continuous Mapping Theorem to the criterion function in Equation 8, we obtain that $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in probability to $Q(\boldsymbol{\beta}, \mathcal{D})$ pointwise for every $\boldsymbol{\beta} \in \mathbb{R}^b$ as $k \to \infty$.

Now in order to invoke Lemma 2, we first define $\mathbb{B}$ to be an open convex subset of $\mathbb{R}^b$ such that the domain of our parameter $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| = 1\} \subset \mathbb{B}$. For instance, we can let $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| < 1+\gamma\}$, for some constant γ. By Lemma 3, Q is convex over $\mathbb{B}$. Hence, by Lemma 2, pointwise convergence of $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ to $Q(\boldsymbol{\beta}, \mathcal{D})$ for every $\boldsymbol{\beta} \in \mathbb{B}$ implies that $\sup_{\boldsymbol{\beta} \in \mathbb{B}} \left| Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D}) \right| \xrightarrow{p} 0$, as $k \to \infty$. Since:

$$\sup_{\boldsymbol{\beta} \in \Theta \subset \mathbb{B}} \left| Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D}) \right| \le \sup_{\boldsymbol{\beta} \in \mathbb{B}} \left| Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D}) \right| \xrightarrow{p} 0 \tag{13}$$

It follows that the criterion function $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ defined over $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| = 1\}$ converges uniformly to $Q(\boldsymbol{\beta}, \mathcal{D})$. That is, $\sup_{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\|=1} \left| Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D}) \right| \xrightarrow{p} 0$, as $k \to \infty$.

□

Assumption 1 (Existence of an identified set Θ^*): Denote Θ as the domain of the parameters, i.e. $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| = 1\}$. There exists a set $\Theta^* \subset \Theta$ such that $\sup_{\boldsymbol{\beta} \in \Theta^*} Q(\boldsymbol{\beta}, \mathcal{D}) = \inf_{\boldsymbol{\beta} \in \Theta} Q(\boldsymbol{\beta}, \mathcal{D})$, and $\forall \nu > 0$, $\inf_{\boldsymbol{\beta} \notin B(\Theta^*, \nu)} Q(\boldsymbol{\beta}, \mathcal{D}) > \sup_{\boldsymbol{\beta} \in \Theta^*} Q(\boldsymbol{\beta}, \mathcal{D})$, where $B(\Theta^*, \nu)$ denotes a union of open balls of radius ν each centered on each element of Θ^*.

Theorem 2 (Convergence of random projection estimator): Suppose that Assumption 1 holds. For any given data $\mathcal{D}$, the random projection estimator $\hat{\Theta}_k = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^b : ||\boldsymbol{\beta}||=1} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in half-Hausdorff distance to the identified set Θ^* as k grows, i.e. $\sup_{\boldsymbol{\beta} \in \hat{\Theta}_k} \inf_{\boldsymbol{\beta}' \in \Theta^*} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \xrightarrow{p} 0$ as $k \to \infty$.

Proof of Theorem 2: To prove this theorem, we need Assumption 1 and Theorem 1. The proof follows Chernozhukov et al. (2007) closely.

Let $\hat{\Theta}_k$ be the random projection estimator, i.e. $\hat{\Theta}_k = \text{argmin}_{\boldsymbol{\beta} \in \Theta} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}})$. Note that $\hat{\Theta}_k$ is random due only to the randomness in random projections. From Assumption 1, we have that:

for all $\nu > 0, \exists \delta > 0$ such that if $\beta \notin B(\Theta^*, \nu)$ and $\beta_0 \in \Theta^*$ then $Q(\beta, \mathcal{D}) - Q(\beta_0, \mathcal{D}) \geq \delta > 0$. Hence for a given $\hat{\beta} \in \hat{\Theta}_k$ and $\beta_0 \in \Theta^*$, we have:

$$\begin{aligned}
\Pr\left(\hat{\beta} \notin B(\Theta^*, \nu)\right) &\leq \Pr\left(Q(\hat{\beta}, \mathcal{D}) - Q(\beta_0, \mathcal{D}) \geq \delta\right) \\
\Pr\Big(\inf_{\beta \in \Theta^*} \|\hat{\beta} - \beta\| > \nu\Big) &\leq \Pr\left(Q(\hat{\beta}, \mathcal{D}) - Q(\beta_0, \mathcal{D}) \geq \delta\right) \\
\Pr\Big(\sup_{\hat{\beta} \in \hat{\Theta}_k} \inf_{\beta \in \Theta^*} \|\hat{\beta} - \beta\| > \nu\Big) &\leq \Pr\Big(\sup_{\hat{\beta} \in \hat{\Theta}_k} Q(\hat{\beta}, \mathcal{D}) - Q(\beta_0, \mathcal{D}) \geq \delta\Big) \\
&= \Pr\Big(\sup_{\hat{\beta} \in \hat{\Theta}_k} Q(\hat{\beta}, \mathcal{D}) - Q(\hat{\beta}, \tilde{\mathcal{D}}) + Q(\hat{\beta}, \tilde{\mathcal{D}}) - Q(\theta_0, \mathcal{D}) \geq \delta\Big) \\
&\leq \Pr\Big(\sup_{\hat{\beta} \in \hat{\Theta}_k} Q(\hat{\beta}, \mathcal{D}) - Q(\hat{\beta}, \tilde{\mathcal{D}}) + Q(\theta_0, \tilde{\mathcal{D}}) - Q(\theta_0, \mathcal{D}) \geq \delta\Big) \\
&\leq \Pr\Big(2 \sup_{\beta \in \Theta} \left|Q(\beta, \tilde{\mathcal{D}}) - Q(\beta, \mathcal{D})\right| \geq \delta\Big) \\
&\xrightarrow{p} 0, \text{ as } k \to \infty \text{ by Theorem 1}
\end{aligned}$$

Therefore all elements of the random projection estimator (i.e. the argmin set of $Q(\beta, \tilde{\mathcal{D}})$) converges to some element of Θ^*, and we have $\sup_{\hat{\beta} \in \hat{\boldsymbol{\beta}}} \inf_{\beta \in \Theta^*} \|\hat{\beta} - \beta\| \xrightarrow{p} 0$. This is convergence in half-Hausdorff distance. Full convergence is not guaranteed, as the random projection estimator might not include enough elements to converge to every element in the identified set. It is possible to strengthen the notion of half-Hausdorff convergence to full Hausdorff convergence following the additional assumptions in Chernozhukov et al. (2007).

□

A.3. Additional convergence result

Assumption 2. As $d \to \infty$, the (deterministic) sequence of data $\mathcal{D}_d = \{(\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(n)}, \boldsymbol{p}^{(n)})\}$ satisfies the following two assumptions. (i) Let $\boldsymbol{X}$ be any vector of covariates in $\mathcal{D}_d$, then $\frac{1}{d}\sum_{j=1}^{d}(X_j)^4$ exist and is bounded as $d \to \infty$. Secondly, without loss of generality, assume that for all vectors of covariates $\boldsymbol{X}$ in the data $\mathcal{D}_d$, $\sum_{j=1}^{d} X_j^2 = \|\boldsymbol{X}\|^2 = O(1)$ as $d \to \infty$. This part is without loss of generality as the cardinality of utilities can be rescaled.

As before, the only source of randomness is in the random projection. Accordingly, the sequence of data $\mathcal{D}_d$ as d grows is deterministic.

Theorem 3. *Suppose that as $d \to \infty$, the sequence of data $\mathcal{D}_d$ satisfies Assumptions 1 and 2. Let R be a $k \times d$ sparse random projection matrix with parameter $s = O(\sqrt{d})$. Denote $\tilde{\mathcal{D}}_k = R\mathcal{D}_d$ as the projected-down data. The random projection estimator $\tilde{\Theta}_k = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^b : \|\boldsymbol{\beta}\| = 1} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in half-Hausdorff distance to the identified set Θ^*, i.e. $\sup_{\boldsymbol{\beta} \in \tilde{\Theta}_k} \inf_{\boldsymbol{\beta}' \in \Theta^*} \|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \xrightarrow{p} 0$ for any sequence of k, d such that $k, d \to \infty$ and $d > k$.*

Proof. Let $\boldsymbol{u}^{(i)} \equiv \boldsymbol{X}^{(i)}\boldsymbol{\beta}$ be the d-dimensional vector of utilities that market i derives from each of the d choices (before realization of shocks), and let $\boldsymbol{p}^{(i)}$ be the corresponding observed choice probabilities for market i in the data $\mathcal{D}_d$. For any $\boldsymbol{\beta}$, and any pair of markets $(a,b) \in \{1,\ldots,n\}^2$, we have from Equation 7:

$$\mathrm{Var}(\|\tilde{\boldsymbol{u}}^{(a)} - \tilde{\boldsymbol{p}}^{(b)}\|^2) = \frac{1}{k}\Big(2\|\boldsymbol{u}^{(a)} - \boldsymbol{p}^{(b)}\|^4 + (s-3)\sum_{j=1}^{d}(u_j^{(a)} - p_j^{(b)})^4\Big) \tag{14}$$

where $(\tilde{\boldsymbol{u}}^{(a)}, \tilde{\boldsymbol{p}}^{(b)})$ denotes the projected-down vector $(R\boldsymbol{u}^{(a)}, R\boldsymbol{p}^{(b)})$, and R is a $k \times d$ sparse random projection matrix with parameter $s = O(\sqrt{d})$.

From Lemma 2 of Li et al. (2006), they show that when $\frac{1}{d}\sum_{j=1}^{d} u_i$ and $\frac{1}{d}\sum_{j=1}^{d} v_i$ exist in the limit as $d \to \infty$, then:

$$\begin{aligned}\mathrm{Var}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) &= \frac{1}{k}\Big(2\|\boldsymbol{u} - \boldsymbol{v}\|^4 + (s-3)\sum_{j=1}^{d}(u_j - v_j)^4\Big) &(15)\\ &= O\left(\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{k}\right) &(16)\end{aligned}$$

Applying Lemma 2 of Li et al. (2006) to Equation 14, if the limit of $\frac{1}{d}\sum_{j=1}^{d}(u_j^{(a)})^4$ exists and is bounded as d grows, then the first term in Equation 14 dominates the second term. Note that $\boldsymbol{p}^{(b)}$ is a vector of choice probabilities whose length is always 1, and so $\frac{1}{d}\sum_{j=1}^{d}(p_j^{(b)})^4$ is bounded as $d \to \infty$.

Plugging in $u_j^{(a)} = \sum_{t=1}^{p} \beta_t X_{j,t}^{(a)}$, we then have $\frac{1}{d}\sum_{j=1}^{d}(u_j^{(a)})^4 = \frac{1}{d}\sum_{j=1}^{d}\left(\sum_{t=1}^{p}\beta_t X_{j,t}^{(a)}\right)^4$. Note that $|\beta_t| < 1$ per the normalization of parameters using $\|\boldsymbol{\beta}\| = 1$. Now a sufficient condition for $\frac{1}{d}\sum_{j=1}^{d}\left(\sum_{t=1}^{p}\beta_t X_{j,t}^{(a)}\right)^4$ to exist in the limit is that $\frac{1}{d}\sum_{j=1}^{d}(X_{j,t}^{(a)})^4$ exists for all t, as stipulated in Assumption (2). (By the Jensen's inequality, if the fourth moment exists, then all lower moments exist, and by the Cauchy–Schwarz inequality, if $\mathbb{E}[X^4]$ and $\mathbb{E}[Y^4]$ exist, then $\mathbb{E}[X^2Y^2]$, $\mathbb{E}[XY^2]$ and so on also exist.)

Having established that the first term in Equation 14 dominates, we now examine the first term. If $\|\boldsymbol{u}^{(a)}\| = O(1)$, then $\|\boldsymbol{u}^{(a)} - \boldsymbol{p}^{(b)}\|^4 = O(1)$ since $\boldsymbol{p}^{(b)}$ is a vector of choice probabilities and $\|\boldsymbol{p}^{(b)}\|$ is bounded for all d. A sufficient condition for $\|\boldsymbol{u}^{(a)}\| \equiv \|\boldsymbol{X}^{(a)}\boldsymbol{\beta}\| = O(1)$ is the following: for all columns $\boldsymbol{X}$ of $\boldsymbol{X}^{(a)}$, we have $\|\boldsymbol{X}\|^2 = O(1)$ as d grows. This is maintained by Assumption 2.

Therefore we can rewrite Equation 14 as $\mathrm{Var}(\|\tilde{\boldsymbol{u}}^{(a)} - \tilde{\boldsymbol{p}}^{(b)}\|^2) = O\left(\frac{1}{k}\right)$. Now this bound of the variance depends only on k and not on d. Therefore as long as $k, d \to \infty$, this variance goes to zero. Hence the criterion function $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges pointwise to $Q(\boldsymbol{\beta}, \mathcal{D})$ as we let both d and k grow. The rest of the proof now follows from Theorems 1 and 2.

□

A.4. Additional tables

Design	mean LB (s.d.)	mean UB (s.d.)	25th LB, 75th UB
$d = 100, k = 10$	2.3459 (0.2417)	2.3459 (0.2417)	[2.1777, 2.5076]
$d = 500, k = 100$	2.2701 (0.2582)	2.3714 (0.2832)	[2.1306, 2.6018]
$d = 1000, k = 100$	2.4001 (0.2824)	2.4001 (0.2824)	[2.2248, 2.6018]
$d = 5000, k = 100$	2.3766 (0.3054)	2.3766 (0.3054)	[2.1306, 2.6018]
$d = 5000, k = 500$	2.2262 (0.3295)	2.4906 (0.3439)	[1.9892, 2.7667]

Table 9: Random projection estimator with dense random projections, $s = 1$. Replicated 100 times using independently realized random projection matrices. The true value of θ is 2.3562.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, *66*(4), 671–687.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Berry, S. T. & Haile, P. A. (2014). Identification in differentiated products markets using market level data. *Econometrica*, *82*(5), 1749–1797.

Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models1. *Econometrica*, *75*(5), 1243–1284.

Chevalier, J. A., Kashyap, A. K., & Rossi, P. E. (2003). Why don't prices rise during periods of peak demand? evidence from scanner data. *American Economic Review*, *93*(1), 15–37.

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016). Editorialmarketing science and big data.

Chiong, K., Galichon, A., & Shum, M. (2016). Duality in dynamic discrete choice models. *Quantitative Economics*, *7*, 83–115.

Ciliberto, F. & Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, *77*(6), 1791–1828.

Coibion, O., Gorodnichenko, Y., & Hong, G. H. (2015). The cyclicality of sales, regular and effective prices: Business cycle and policy implications. *American Economic Review*, *105*(3), 993–1029.

Dasgupta, S. & Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, *22*(1), 60–65.

Davis, D. R., Dingel, J. I., Monras, J., & Morales, E. (2016). How segregated is urban consumption? Technical report, Columbia University.

Fosgerau, M. & De Palma, A. (2015). Generalized entropy models. Technical report. working paper, Technical University of Denmark.

Fosgerau, M. & McFadden, D. (2012). A theory of the perturbed consumer with general budgets. Technical report, NBER Working Paper No. 17953.

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *RAND Journal of Economics*, 1002–1019.

Fox, J. T. & Bajari, P. (2013). Measuring the efficiency of an fcc spectrum auction. *American Economic Journal: Microeconomics*, *5*(1), 100–146.

Gandhi, A., Lu, Z., & Shi, X. (2013). Estimating demand for differentiated products with error in market shares. Technical report, University of Wisconsin-Madison.

Gentzkow, M., Shapiro, J., & Taddy, M. (2016). Measuring polarization in high-dimensional data: Method and application to congressional speech. Technical report, Stanford University.

Gillen, B. J., Montero, S., Moon, H. R., & Shum, M. (2015). Blp-lasso for aggregate discrete choice models of elections with rich demographic covariates. *USC-INET Research Paper*, (15-27).

Goeree, J. K., Holt, C. A., & Palfrey, T. R. (2005). Regular quantal response equilibrium. *Experimental Economics*, *8*(4), 347–367.

Haile, P. A., Hortaçsu, A., & Kosenok, G. (2008). On the empirical content of quantal response equilibrium. *American Economic Review*, *98*(1), 180–200.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, *35*(2), 303–316.

Hausman, J. & McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 1219–1240.

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

Heffetz, O. & Ligett, K. (2014). Privacy and data-based research. *Journal of Economic Perspectives*, *28*(2), 75–98.

Huang, D. & Luo, L. (2016). Consumer preference elicitation of complex products using fuzzy support vector machine active learning. *Marketing Science*, *35*(3), 445–464.

Ichimura, H. & Lee, L.-F. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*, (pp. 3–49).

Jacobs, B. J., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, *35*(3), 389–404.

Keane, M. & Wasi, N. (2012). Estimation of discrete choice models with many alternatives using random subsets of the full choice set: With an application to demand for frozen pizza. Technical report, Oxford University.

Lee, L.-F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, *65*(2), 381–428.

Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 287–296). ACM.

Li, P., Mitzenmacher, M., & Shrivastava, A. (2014). Coding for random projections. In *ICML*, (pp. 676–684).

Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, *35*(3), 363–388.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, *3*(3), 205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, *27*(3), 313–333.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In *Frontiers in Econometrics*, ed. P. Zarembka.(New York: Academic Press).

McFadden, D. (1978). Modelling the choice of residential location. Technical report, Institute of Transportation Studies, University of California-Berkeley.

McFadden, D. (1981). Econometric models of probabilistic choice. In C. Manski & D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press.

Melo, E., Pogorelskiy, K., & Shum, M. (2015). Testing the quantal response hypothesis. Technical report, California Institute of Technology.

Ng, S. (2015). Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, Columbia University.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, *7*(02), 186–199.

Powell, J. L. & Ruud, P. A. (2008). Simple estimators for semiparametric multinomial choice models. Technical report, University of California, Berkeley.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton university press.

Shi, X., Shum, M., & Song, W. (2016). Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. Technical report, University of Wisconsin-Madison.

Train, K. E., McFadden, D. L., & Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *RAND Journal of Economics*, 109–123.

Vempala, S. (2000). *The Random Projection Method.* American Mathematical Society. Series in Discrete Mathematics and Theoretical Computer Science (DIMACS), Vol. 65.

Villani, C. (2003). *Topics in optimal transportation.* Number 58. American Mathematical Society.