DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

RANDOM PROJECTION ESTIMATION OF DISCRETE-CHOICE MODELS WITH LARGE CHOICE SETS

Khai X. Chiong

Matthew Shum

SOCIAL SCIENCE WORKING PAPER 1416

March 2016

Random Projection Estimation of Discrete-Choice Models with Large Choice Sets

Khai X. Chiong Matthew Shum

Abstract

We introduce *sparse random projection*, an important dimension-reduction tool from machine learning, for the estimation of discrete-choice models with high-dimensional choice sets. Initially, the high-dimensional data are compressed into a lower-dimensional Euclidean space using random projections. Subsequently, estimation proceeds using cyclic 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. The random projection procedure is justified via the Johnson-Lindenstrauss Lemma: – the pairwise distances between data points are preserved during data compression, which we exploit to show convergence of our estimator. The estimator works well in a computational simulation and in a application to a supermarket scanner dataset.

JEL classification numbers: C14, C25, C55

Key words: semiparametric multinomial choice models, random projection, large choice sets, cyclic monotonicity, Johnson-Lindenstrauss Lemma

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

Khai X. Chiong† Matthew Shum‡

1. Introduction

Estimation of discrete-choice models in which consumers face high-dimensional choice sets 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. As far as we are aware, this is the first use of random projection in the econometrics literature on 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,

*First draft: February 29, 2016. This draft: March 2016. We thank Hiroaki Kaido, Michael Leung, Sergio Montero, and participants at the DATALEAD conference (Paris, November 2015) for helpful comments.

†INET & University of Southern California. E-mail: kchiong@usc.edu

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

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 $k << d$) stochastic matrix, resulting in a smaller k-by-l compressed data matrix that is more manageable. Subsequenty, we estimate the discrete-choice model using the compressed data matrix, in place of the original high-dimensional dataset. Specifically in the second step, we estimate the discrete-choice model without needing to specify the distribution of the random utility errors by using inequalities derived from *cyclic monotonicity*: – a generalization of the notion of monotonicity for vector-valued functions which always holds for random-utility discrete-choice models (Rockafellar (1970), Chiong et al. (2016)).

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 easy. For instance, when the dimensionality of the choice set is $d = 5,000$, the random projection matrix consists of roughly 99% zeros, and indeed only 1% of the data matrix is needed or sampled.

We show theoretically that the random projection estimator converges to the unprojected estimator, as k grows large. We utilize results from the machine learning literature, which show that random projection enables embeddings of points from high-dimensional into low-dimensional Euclidean space with high probability, and hence we can consistently recover the original estimates from the compressed dataset. In the simulation, even with small and moderate k, we show that the noise introduced by random projection is reasonably small. In summary, k controls the trade-off between using a small/tractable dataset for estimation, and error in estimation.

As an application of our procedures, we estimate a 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]For example, Ng (2015) analyzes terabytes of scanner data that required an amount of RAM that was beyond the budget of most researchers

1.1. Related Literature

The difficulties with 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 et al. (1987) (see also Davis et al. (2016)). This sampling approach depends crucially on the multinomial logit assumption on the errors, and particularly on the independence of the errors across items in the large choice set. Relatedly, Gentzkow et al. (2016) use a Poisson approximation to enable parallel computing a multinomial logit model of legislators' choices among hundreds of thousands of phrases.

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 (cf. Manski (1975, 1985)) to estimate semiparametric multinomial choice models with large choice sets but using only a subset of the choices.[2] Identification relies on a "rank-order" assumption, which is satisfied by exchangeability of the joint error distribution (we discuss this in more detail below). The rank-order property is an implication of the Independence of Irrelevant Alternatives (IIA) property, and hence can be considered as a generalized version of IIA. In contrast, our cyclic 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.[3] Particularly, our approach accommodates models with error structures in the generalized extreme value family (ie. nested logit models), 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.

[2] Fox and 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.

[3] Besides Fox (2007), the literature on semiparametric multinomial choice models is quite small, and includes the multiple-index approach of Ichimura and Lee (1991) and Lee (1995), and a pairwise-differencing approach in Powell and 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.

Indeed, Fox's rank-order property and the cyclic 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 cyclic monotonicity restricts the dependence of the utility shocks across different markets (or choice scenarios).[4]

The ideas of random projection were popularized in the Machine Learning literature on dimensionality reduction (Achlioptas (2003); Dasgupta and Gupta (2003); 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.

However, in many high-dimensional econometric models only the lengths and inner products 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.[5]

Section 2 presents our semiparametric discrete-choice modeling framework, and the moment inequalities derived from cyclic 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

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

[5] See Belloni et al. (2012), Belloni et al. (2014), and Gillen et al. (2015). Both LASSO and 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.

converges to the full-sample estimator as the dimension of projection increases. Section 4 contains results from simulation examples, demonstrating the well-working of random projection 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 in which 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.[6] Specifically, an agent chooses from among $\mathcal{C} = [1, \ldots, d]$ alternatives or choices. High-dimensionality here refers to a large value of d. The utility that the agent derives from choice j is $\boldsymbol{X}_j\boldsymbol{\beta} + \epsilon_j$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b)' \in \mathbb{R}^b$ are unknown parameters, and $\boldsymbol{X}_j$ is a $1 \times b$ vector of covariates specific to choice j. Here, ϵ_j is a utility shock, encompassing unobservables which affect the agent's utility from the j-th choice.

Let $u_j \equiv \boldsymbol{X}_j\boldsymbol{\beta}$ denote the deterministic part of utility that the agent derives from choice j, and 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 the agent chooses j is $p_j(\boldsymbol{u}) = \Pr(u_j + \epsilon_j \geq \max_{k \neq j}\{u_k + \epsilon_k\})$. 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 $\boldsymbol{\epsilon} \equiv (\epsilon_1, \ldots, \epsilon_d)'$ are distributed independently of $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$, but otherwise allow it to follow an unknown joint distribution that can be arbitrarily correlated among different choices j. 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* **cyclic monotonicity**.

Definition 1 (Cyclic 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 $(\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^L, \boldsymbol{u}^1)$. The

[6] Virtually all the existing papers on semiparametric multinomial choices use similar setups (Fox (2007), Ichimura and Lee (1991), Lee (1995), Powell and Ruud (2008)).

function $\boldsymbol{p}$ is cyclic monotone with respect to the cycle $(\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^L, \boldsymbol{u}^1)$ if and only if

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

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

Proposition 1 arises from underlying convexity properties of the discrete-choice problem. We refer to Chiong et al. (2016) and Shi et al. (2016) for the full details. Briefly, the independence of $\boldsymbol{\epsilon}$ and $\boldsymbol{X}$ implies that the *social surplus function* of the discrete choice model, defined as,

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

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 $\boldsymbol{u}$;[7] 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 cyclic monotonicity, and hence satisfies the CM-inequalities in (1) above. (In fact, any function that satisfies cyclic monotonicity must be a subgradient of some convex function.) Therefore, cyclic monotonicity is the appropriate vector generalization of the fact that the slope of a scalar-valued convex function is monotone increasing.

2.1. Inequalities for estimation

Following Shi et al. (2016), we use the cyclic monotonic inequalities in (1) to estimate the parameters $\boldsymbol{\beta}$.[8] Suppose we observe the aggregate behavior of many independent agents across n different markets.[9] Our dataset consists of $\mathcal{D} = \left((\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(n)}, \boldsymbol{p}^{(n)})\right)$,

[7] 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.

[8] See also Melo et al. (2015) for an application of cyclic monotonicity for testing game-theoretic models of stochastic choice.

[9] Throughout this paper, we assume the researcher has access to such aggregate data, in which the market-level choice probabilities (or market shares) are directly observed. Such data structures arise often in aggregate demand models in empirical industrial organization (eg. Berry and Haile (2014), Gandhi et al. (2013)). We do not consider the application to individual-level choice data in this paper.

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

$$\sum_{l=1}^{L}(\boldsymbol{X}^{(a_{l+1})}\boldsymbol{\beta} - \boldsymbol{X}^{(a_l)}\boldsymbol{\beta}) \cdot \boldsymbol{p}^{(a_l)} \leq 0, \quad \text{for all cycles } (a_l)_{l=1}^{L+1} \text{ in data } \mathcal{D}, L \geq 2 \tag{3}$$

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 cyclic monotonicity inequalities in (3), 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 cyclic monotonicity inequalities:

$$Q(\boldsymbol{\beta}) = \sum_{\text{all cycles in data } \mathcal{D}; L \geq 2} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})}\boldsymbol{\beta} - \boldsymbol{X}^{(a_l)}\boldsymbol{\beta}\right) \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 \tag{4}$$

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{B}: ||\boldsymbol{\beta}|| = 1}{\operatorname{argmin}} \; Q(\boldsymbol{\beta}).$$

The parameter space $\mathbb{B}$ is defined to be a convex subset of $\mathbb{R}^b$. 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, for instance, in Han (1987) and Hausman et al. (1998). Shi et al. (2016) shows that the criterion function above delivers consistent interval estimates of the identified set of parameters under the assumption that the covariates are exogenous. The criterion function here is convex, and the global minimum can be found using subgradient descent (since it is not differentiable everywhere).[10]

However for reasons discussed earlier, high-dimensional choice sets posed particular challenges for semi-parametric estimation. Next, we describe how random projection can help reduce the dimensionality of our problem.

[10]Because the cyclic monotonicity inequalities involve differences in $\boldsymbol{X}\boldsymbol{\beta}$, no constant terms need be included in the model, as it would simply difference out across markets. Similarly, any outside good with mean utility normalized to zero would also drop out of the cyclic monotonicity inequalities.

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 compressed 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 (4) is computed using only the compressed data $\tilde{\mathcal{D}}$.

A random projection matrix R, is a k-by-d matrix (with $k << d$) such that each entry $R_{i,j}$ is distributed i.i.d according to $\frac{1}{\sqrt{k}}F$, where F is any mean zero distribution. For any d-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, premultiplication by R yields the random reduced-dimensional $(k \times 1)$ vectors $R\boldsymbol{u}$ and $R\boldsymbol{v}$; thus, $R\boldsymbol{u}$ and $R\boldsymbol{v}$ are the random projections of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively.

By construction, a random projection matrix R has the property that, given two high-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, the squared Euclidean distance between the two projected-down vectors $\|R\boldsymbol{u} - R\boldsymbol{v}\|^2$ is a random variable with mean equal to $\|\boldsymbol{u} - \boldsymbol{v}\|^2$, the squared distance between the two original high-dimensional vectors. Essentially, the random projection procedure replaces each high-dimensional vector $\boldsymbol{u}$ with a random lower-dimensional counterpart $\tilde{\boldsymbol{u}} = R\boldsymbol{u}$ the length of which is a mean-preserving spread of the original vector's length.[11]

Most early applications of random projection utilized Gaussian random projection matrices, in which each entry of R is generated independently from standard Gaussian (normal) distributions. However, for computational convenience and simplicity, we focus in this paper on **sparse** random projection matrices, in which many elements will be equal to zero with high probability. Moreover, different choice of probability distributions of $R_{i,j}$ can lead to different variance and error tail bounds of $\|R\boldsymbol{u} - R\boldsymbol{v}\|^2$. Following the work of Li et al. (2006), we introduce a class of sparse random projection matrices that can also be tailored to enhance the efficiency of random projection.

Definition 2 (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

[11] For a detailed discussion, see chap. 1 in Vempala (2000).

according to the following discrete distribution:

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

By choosing a higher s, we produce sparser random projection matrices, but also increase the variance of the projected-down vectors. Li et al. (2006) shows that:

$$\text{Var}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = \frac{1}{k}\Big(2\|\boldsymbol{u} - \boldsymbol{v}\|^4 + (s-3)\sum_{j=1}^{d}(u_j - v_j)^4\Big) \tag{5}$$

By using different values of s, we achieve a trade-off between having a sparser random projection matrix, and one that is more efficient. For instance, if we let $s = 1$, we achieve a dense random projection matrix, but we also achieve the lowest variance. We call this the optimal random projection. If we let $s = 3$, we obtain a variance of $\frac{1}{k}2\|\boldsymbol{u} - \boldsymbol{v}\|^4$, which interestingly, is the same variance achieves by the benchmark Gaussian random projection (each element of the random projection matrix is distributed i.i.d. according to the standard Gaussian, see Achlioptas (2003)).

Since Gaussian random projection is dense and has the same efficiency as the sparse random projection with $s = 3$, the class of random projections proposed in Definition 2 is to preferred in terms of both efficiency and sparsity. Moreover, random uniform numbers are much easier to generate than Gaussian random numbers.

Moreover, Li et al. (2006) argues that in practice, we can use very sparse random projections with little loss in efficiency. In particular, the loss in efficiency for setting a very large value of s (ultra-sparse random projection) vis-à-vis a dense random projection is neglible when d is large, which is precisely the setting where random projection is desired. More concretely, we can set s to be as large as $\sqrt{d}$. We will see in the simulation example that when $d = 5,000$, 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.[12]

[12] 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 $j = 1, \ldots, d$, then the term $\|\boldsymbol{u} - \boldsymbol{v}\|^4$ in the variance formula (Eq. 5) 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).

3.1. Random Projection Estimator

We introduce the random projection estimator. Given the dataset $\mathcal{D} = \{(\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(n)}, \boldsymbol{p}^{(n)})\}$, define the *compressed* dataset by $\tilde{\mathcal{D}}_k = \{(\tilde{\boldsymbol{X}}^{(1)}, \tilde{\boldsymbol{p}}^{(1)}), \ldots, (\tilde{\boldsymbol{X}}^{(n)}, \tilde{\boldsymbol{p}}^{(n)})\}$, where $(\tilde{\boldsymbol{X}}^{(i)}, \tilde{\boldsymbol{p}}^{(i)}) = (R\boldsymbol{X}^{(i)}, R\boldsymbol{p}^{(i)})$ for all markets i, and R being a sparse $k \times d$ random projection matrix as in Definition 2.

Definition 3 (Random projection estimator)**:** The random projection estimator is defined as $\tilde{\boldsymbol{\beta}}_k \in \operatorname{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$, where $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ is the criterion function in Equation (4) in which the input data is $\tilde{\mathcal{D}}_k$. ■

The compressed 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 compressed 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 (4) in which the uncompressed 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 uncompressed data, as $k \to \infty$. Here we provide some intuition and state some preliminary results for this convergence result.

Recall in the previous section that the Euclidean distance between two vectors are preserved in expectation as these vectors are compressed 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 cyclic monotonicity – in terms of Euclidean norms.

Definition 4 (Cyclic 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 cyclic monotone with respect to the cycle $\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots, \boldsymbol{u}^L, \boldsymbol{u}^1$ if and only if

$$\sum_{l=2}^{L+1} \left(\|\boldsymbol{u}^l - \boldsymbol{p}^l\|^2 - \|\boldsymbol{u}^l - \boldsymbol{p}^{l-1}\|^2 \right) \leq 0 \tag{6}$$

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

The inequalities 1 and 6 equivalently defined cyclic monotonicity (see Villani (2003)); a proof is given in the appendix. Therefore, from Definition 4, we can rewrite the estimator in (4) as $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{B}} Q(\boldsymbol{\beta})$ where the criterion function is defined as the sum of squared violations of the cyclic monotonicity inequalities:

$$Q(\boldsymbol{\beta}) = \sum_{\text{all cycles in data } \mathcal{D}; L\geq 2} \left[\sum_{m=2}^{L+1} \left(\|\boldsymbol{X}^{(a_l)}\boldsymbol{\beta} - \boldsymbol{p}^{(a_l)}\|^2 - \|\boldsymbol{X}^{(a_l)}\boldsymbol{\beta} - \boldsymbol{p}^{(a_{l-1})}\|^2 \right) \right]_+^2 \tag{7}$$

To see the 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:

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 $k = \frac{20 \log C}{\delta^2}$. There exists a linear map $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$:*

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

Proofs of the Johnson-Lindenstrauss Lemma can be found in, among others, Dasgupta and Gupta (2003), Vempala (2000). The proof is probabilistic, and demonstrates that, with a non-zero probability, a *random projection* satisfies the error bounds stated in the Lemma. For this reason, the Johnson-Lindenstrauss Lemma has become a term that collectively represents random projection methods, although the implication of the lemma is not directly used.

Remark 1: The feature that the cyclic monotonicity inequalities can be written in terms of Euclidean norms between vectors justifies the application of the Johnson-Lindenstrauss Lemma, and hence random projection, to our estimator, which is based on these inequalities. In contrast, the "rank-order" inequalities, which underlie the maximum score approach to semiparametric multinomial choice estimation,[13] cannot be rewritten in terms in terms of Euclidean norms between data vectors, and hence random projection cannot be used for those inequalities.

Remark 2: The derivation of our estimation approach for discrete-choice models does not imply that all the choice probabilities be strictly positive – that is, zero choice

[13] eg. 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 $u_i > u_j$.

probabilities are allowed for.[14] The possibility of zero choice probabilities is especially important and empirically relevant especially in a setting with large choice sets, as dataset with large choice sets (such as store-level scanner data) often have zero choice probabilities for many products (cf. Gandhi et al. (2013)).

3.2. Convergence

In this section we show that, for any given data $\mathcal{D}$, the random projection estimator computed using the compressed data $\tilde{\mathcal{D}}_k = R \cdot \mathcal{D}$ converges in probability to the corresponding estimator computed using the uncompressed data $\mathcal{D}$, as $k \to \infty$, where k is the number of rows in the random projection matrix R.

Therefore, k **controls the trade-off between tractability and error in estimation.** These results do not depend on d, the dimension of the choice sets (which is also the number of columns of R.) In order to highlight the random [projection aspect of our estimator, we assume that the market shares and other data variables are assumed to be observed without error. Hence, given the original (uncompressed) 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 for results in this section are provided in Appendix C. We first show that the random-projected criterion function converges uniformly to the unprojected criterion function:

Theorem 1 (Uniform convergence of criterion function)**.** *For any given dataset $\mathcal{D}$, we have* $\sup_{\boldsymbol{\beta} \in \mathbb{B}} |Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k) - Q(\boldsymbol{\beta}, \mathcal{D})| \xrightarrow{p} 0$*, 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 *Hausdorff* distance, where the Hausdorff

[14] Specifically, Eq. (2) allows some of the components of the choice probability vector $\boldsymbol{p}(\boldsymbol{u})$ to be zero.

distance is a distance measure between two sets is: $d(X,Y) = \sup_{y\in Y} \inf_{x\in X} \|x-y\| + \sup_{x\in X} \inf_{y\in Y} \|x-y\|$.

Assumption 1 (Existence of identified set Θ^*)**.** For any given data $\mathcal{D}$, we assume that there exists a set Θ^* (that depends on $\mathcal{D}$) 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 hold. For any given data $\mathcal{D}$, the random projection estimator $\tilde{\Theta}_k = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{B}} Q(\boldsymbol{\beta},\tilde{\mathcal{D}}_k)$ converges in half-Hausdorff distance to the identified set Θ^* as k grows, i.e. $\sup_{\boldsymbol{\beta}\in\tilde{\Theta}_k} \inf_{\boldsymbol{\beta}'\in\Theta^*} \|\boldsymbol{\beta}-\boldsymbol{\beta}'\| \xrightarrow{p} 0$ as k grows.*

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. 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. Therefore the market shares and other data variables are assumed to be observed without error.

The main conclusion from this section is that the error introduced by random projection is negligible, even when the reduced dimension k is very small. In the tables below, we see that the random projection method produces interval estimates that are always strictly nested within the identified set which was obtained when the full uncompressed data are used.

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.[15]

[15] 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 $\eta_{l,j}^{(m)} \sim N(0,1)$.

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. 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 $\mathcal{D} = \{(\boldsymbol{X}^{(1)}, \boldsymbol{p}^{(1)}), \ldots, (\boldsymbol{X}^{(n)}, \boldsymbol{p}^{(n)})\}$ for $n = 30$ markets, where $\boldsymbol{p}^{(m)}$ corresponds to the d-by-1 vector of simulated choice probabilities for market m: the j-th row of $\boldsymbol{p}^{(m)}$ is $\boldsymbol{p}_j^{(m)} = \Pr\left(U_j^{(m)} + \epsilon_j^{(m)} > U_{-j}^{(m)} + \epsilon_{-j}^{(m)}\right)$. We then perform random projection on $\mathcal{D}$ to obtain the compressed dataset $\tilde{\mathcal{D}} = \{(\tilde{\boldsymbol{X}}^{(1)}, \tilde{\boldsymbol{p}}^{(1)}), \ldots, (\tilde{\boldsymbol{X}}^{(n)}, \tilde{\boldsymbol{p}}^{(n)})$ More specifically, where for all markets m, $(\tilde{\boldsymbol{X}}^{(m)}, \tilde{\boldsymbol{p}}^{(m)}) = (R\boldsymbol{X}^{(i)}, R\boldsymbol{p}^{(i)})$, where R is a realized $k \times d$ random projection matrix as in Definition 2. Having constructed the compressed dataset, the criterion function in Eq. 4 is used to estimate β. We restrict to cycles of length 2 and 3 in computing Eq. 4; however, we find that even using cycles of length 2 did not change the result in any noticeable way.

The random projection matrix is parameterized by s (see Definition 2). We set $s = 1$, which corresponds to the optimal random projection matrix. In Table 2, we show that sparse random projections ($s = \sqrt{d}$ in Definition 2) perform just as well. Sparse random projections are much faster to perform – for instance when $d = 5000$, we sample less than 2% of the data, as over 98% of the random projection matrix are zeros.

In these tables, 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 last column reports the actual identified set that is computed without using random projections. (In the Appendix, Tables 5 and 6, we see that in all the runs, our approach produces interval estimates that are always strictly nested within the identified sets.)

The results indicate that, in most cases, optimization of the randomly-projected criterion function $Q(\boldsymbol{\beta}, \mathcal{D}_k)$ yields a unique minimum, in contrast to the unprojected criterion

function $Q(\boldsymbol{\beta}, \mathcal{D})$, which is minimized at an interval. For instance, in the fourth row of Table 1 (when compressing from $d = 5000$ to $k = 100$), we see that the true identified set for this specification, computed using the unprojected data, is $[1.2038, 3.5914]$, but the projected criterion function is always uniquely minimized (across all 100 replications). Moreover the average point estimate for θ is equal to 2.3766, where the true value is 2.3562. This is unsurprising, 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.

Table 1: Random projection estimator with optimal random projections, $s = 1$

Design	mean LB (s.d.)	mean UB (s.d.)	25th LB, 75th UB	True id set
$d = 100, k = 10$	2.3459 (0.2417)	2.3459 (0.2417)	[2.1777, 2.5076]	[1.4237, 3.2144]
$d = 500, k = 100$	2.2701 (0.2582)	2.3714 (0.2832)	[2.1306, 2.6018]	[1.2352, 3.4343]
$d = 1000, k = 100$	2.4001 (0.2824)	2.4001 (0.2824)	[2.2248, 2.6018]	[1.1410, 3.4972]
$d = 5000, k = 100$	2.3766 (0.3054)	2.3766 (0.3054)	[2.1306, 2.6018]	[1.2038, 3.5914]
$d = 5000, k = 500$	2.2262 (0.3295)	2.4906 (0.3439)	[1.9892, 2.7667]	[1.2038, 3.5914]

Replicated 100 times using independently realized random projection matrices. The true value of θ is 2.3562. Right-most column reports the interval of points that minimized the unprojected criterion function.

Table 2: Random projection estimator with sparse random projections, $s = \sqrt{d}$

Design	mean LB (s.d.)	mean UB (s.d.)	25th LB, 75th UB	True id set
$d = 100, k = 10$	2.3073 (0.2785)	2.3073 (0.2785)	[2.1306, 2.5076]	[1.4237, 3.2144]
$d = 500, k = 100$	2.2545 (0.2457)	2.3473 (0.2415)	[2.0363, 2.5076]	[1.2352, 3.4343]
$d = 1000, k = 100$	2.3332 (0.2530)	2.3398 (0.2574)	[2.1777, 2.5076]	[1.1410, 3.4972]
$d = 5000, k = 100$	2.3671 (0.3144)	2.3671 (0.3144)	[2.1777, 2.5547]	[1.2038, 3.5914]
$d = 5000, k = 500$	2.3228 (0.3353)	2.5335 (0.3119)	[2.1306, 2.7667]	[1.2038, 3.5914]

Replicated 100 times using independently realized **sparse** random projection matrices (where $s = \sqrt{d}$ in Definition 2). The true value of θ is 2.3562. Right-most column reports the interval of points that minimized the unprojected criterion function.

5. Empirical Application: a discrete-choice 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.[16] 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, at each Dominick's store. 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 et al. (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, and sensitive to different researchers' specific assumption 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.[17]

In this context, a benefit of our semiparametric is that we are *agnostic* as to the joint distribution of utility shocks; hence our approach accommodate 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

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

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

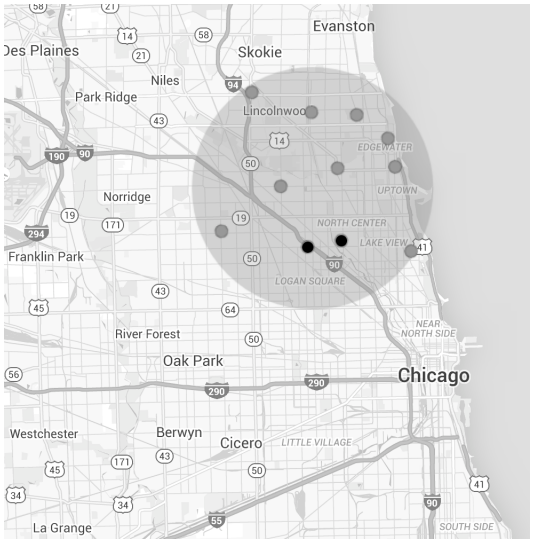

Figure 1: Location of the 11 Dominick's branches as indicated by spots.

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

interval between October 3rd 1996 to April 30th 1997, which is the last recorded date. We include sales at eleven Dominicks supermarkets in north-central Chicago, as illustrated in Figure 1. Among these eleven supermarkets, most are classified as premium-tier stores, while two are medium-tier stores (distinguished by dark black spots in Figure 1); stores in different tiers sell different ranges of products.

Our store and brand choice model consists of $d = 3059$ choices, each corresponds to a unique store and UPC combination. We also define an outside option, for a total of $d = 3060$ choices.[19]The summary statistics for our data sample are in Table 3.

Table 4 presents the estimation results. As in the simulation results above, we ran 100 independent random projections, and thus obtained 100 sets of parameter estimates, for each model specification. The results reported in Table 4 are therefore summary statistics of the estimates for each parameter. 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.[20]

[18]Stores in the same tier share similar product selection, and also pricing to a certain extent.

[19]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 $1 - \sum_{j=1}^{3093} p_{ij}$.

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

	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.[18]	2 out of 11 stores
d	Number of store-upc	3059

Table 3: Summary statistics

Total number of observations is 45885, i.e. number of store-upc=3059 times number of markets (two-week periods)=15.

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.[21] In addition, the interaction between *price* and *holiday* is strongly negative across specifications, indicating that households are more price-sensitive during the holiday season. For the magnitude of this effect, consider a soft drink product priced initially at \$1.00 with no promotion. The median parameter estimates for Specification (C) suggest that during the holiday 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.[22]

We also obtain a positive sign on *bonus*, and the negative sign on the interaction *price* $\times$ *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 (ie. make the price coefficient

[21] cf. Chevalier et al. (2003).

[22] $-0.77\alpha = 0.0661 - (0.77 + 0.36)\alpha(1 + 0.27)$, where $\alpha = -0.1161$ equals 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. (5).)

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]
	$k = 300$			
	Cycles of length 2 & 3			

Table 4: 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 (Eq. 4).

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 time-invariant variable and, apparently here, interacting it with price still does not result in enough variation for reliable estimation.

6. Concluding remarks

In this paper we consider the use of random projection – an important tool for dimension-reduction from machine learning – for estimating multinomial-choice models with large choice sets, a model which arises in many empirical applications. Unlike many recent ap-

plications of machine learning in econometrics, dimension-reduction here is not required for selecting amongst high-dimensional covariates, but rather for reducing the inherent high-dimensionality of the model (ie. reducing the size of agents' choice sets).

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 results in machine learning, which shows that the pairwise distances between data points are preserved during data compression. As we show, in practice the random projection can be very sparse, in the sense that only a small fraction (1%) of the dataset is used in constructing the projection. In the second step, estimation proceeds using the cyclic 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 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.[23] We are currently exploring additional applications of random projection for econometric settings in which privacy may be an issue.

[23] cf. Heffetz and Ligett (2014).

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

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

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

Chevalier, J. A., Kashyap, A. K., and 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.

Chiong, K., Galichon, A., and Shum, M. (2016). Duality in dynamic discrete choice models. *Quantitative Economics*. forthcoming.

Coibion, O., Gorodnichenko, Y., and 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. and 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., and Morales, E. (2016). How segregated is urban consumption? Technical report, Columbia University.

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

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

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

Gentzkow, M., Shapiro, J., and 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., and 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., and Palfrey, T. R. (2005). Regular quantal response equilibrium. *Experimental Economics*, 8(4):347–367.

Haile, P. A., Hortaçsu, A., and 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. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, pages 1219–1240.

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

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

Ichimura, H. and 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*, pages 3–49.

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

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 Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press.

Melo, E., Pogorelskiy, K., and 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. and 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., and 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., and 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*, pages 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.

Appendix A Additional Tables and Figures

Design	min LB, max UB	True id set
$d = 100, k = 10$	[1.8007, 3.3087]	[1.4237, 3.2144]
$d = 500, k = 100$	[1.7536, 2.9317]	[1.2352, 3.4343]
$d = 1000, k = 100$	[1.6593, 2.9317]	[1.1410, 3.4972]
$d = 5000, k = 100$	[1.6593, 3.1202]	[1.2038, 3.5914]
$d = 5000, k = 500$	[1.6593, 3.1202]	[1.2038, 3.5914]

Table 5: Random projection estimator with optimal random projections, $s = 1$. Replicated 100 times using independently realized random projection matrices. The true value of θ is 2.3562. Identified set is the interval of points that minimized the unprojected criterion function.

Design	min LB, max UB	True id set
$d = 100, k = 10$	[1.4237, 2.9788]	[1.4237, 3.2144]
$d = 500, k = 100$	[1.7536, 2.9788]	[1.2352, 3.4343]
$d = 1000, k = 100$	[1.6122, 3.0259]	[1.1410, 3.4972]
$d = 5000, k = 100$	[1.4237, 3.3558]	[1.2038, 3.5914]
$d = 5000, k = 500$	[1.6593, 3.0259]	[1.2038, 3.5914]

Table 6: 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 2). The true value of θ is 2.3562. Identified set is the interval of points that minimized the unprojected criterion function.

Appendix B Equivalence of alternative representation of cyclic monotonicity

Here we show the equivalence of Eqs. (1) and (6), as two alternative statements of the cyclic monotonicity inequalities. We begin with the second statement (6). 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

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

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

□

Appendix C Proof of Theorems in Section 3.2

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. 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 given dataset $\mathcal{D}$, where $\mathbb{B}$ is an open convex subset of $\mathbb{R}^b$.*

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

$$
\begin{aligned}
& Q(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}') \\
= & \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \left(\lambda\boldsymbol{\beta} + (1-\lambda)\boldsymbol{\beta}' \right) \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 \\
= & \sum_{\text{all cycles in data } \mathcal{D}} \left[\lambda \sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta} \cdot \boldsymbol{p}^{(a_l)} + (1-\lambda) \sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta}' \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 \\
\leq & \sum_{\text{all cycles in data } \mathcal{D}} \left\{ \lambda \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta} \cdot \boldsymbol{p}^{(a_l)} \right]_+ + (1-\lambda) \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta}' \cdot \boldsymbol{p}^{(a_l)} \right]_+ \right\}^2 \qquad (8) \\
\leq & \lambda \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta} \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 + \\
& (1-\lambda) \sum_{\text{all cycles in data } \mathcal{D}} \left[\sum_{l=1}^{L} \left(\boldsymbol{X}^{(a_{l+1})} - \boldsymbol{X}^{(a_l)} \right) \boldsymbol{\beta}' \cdot \boldsymbol{p}^{(a_l)} \right]_+^2 \qquad (9) \\
= & \lambda Q(\boldsymbol{\beta}) + (1-\lambda) Q(\boldsymbol{\beta}')
\end{aligned}
$$

Inequality 8 above is due to the fact that $\max\{x,0\} + \max\{y,0\} \geq \max\{x+y,0\}$ for all $x, y \in \mathbb{R}$. Inequality 9 holds from the convexity of the function $f(x) = x^2$. □

Proof of Theorem 1: Recall from Eq. (5) 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 2, we have:

$$\mathbb{E}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = \|\boldsymbol{u} - \boldsymbol{v}\|^2 \qquad (10)$$

$$\mathrm{Var}(\|R\boldsymbol{u} - R\boldsymbol{v}\|^2) = O\left(\frac{1}{k}\right) \qquad (11)$$

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 any 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 versions of $\boldsymbol{X}$ and $\boldsymbol{p}$. Applying the Continuous Mapping Theorem to the criterion function in Eq. 7, we obtain that $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges in probability to $Q(\boldsymbol{\beta}, \mathcal{D})$ pointwise for every $\boldsymbol{\beta}$ as $k \to \infty$.

By Lemma 3, the criterion function $Q(\boldsymbol{\beta}, \mathcal{D})$ is convex in $\boldsymbol{\beta} \in \mathbb{B}$ for any given data $\mathcal{D}$, where $\mathbb{B}$ is an open convex subset of $\mathbb{R}^b$. Therefore, we can immediately invoke the Convexity Lemma to show that pointwise convergence of the Q function implies that $Q(\boldsymbol{\beta}, \tilde{\mathcal{D}}_k)$ converges uniformly to $Q(\boldsymbol{\beta}, \mathcal{D})$. □

Proof of Theorem 2: The result follows readily from Assumption 1 and Theorem 1, by invoking Chernozhukov et al. (2007). The key is in recognizing that (i) in our random finite-sampled criterion function, the randomness stems from the k-by-d random projection matrix,

(ii) the deterministic limiting criterion function here is defined to be the criterion function computed without random projection, taking the full dataset as given. We can then strengthen the notion of half-Hausdorff convergence to full Hausdorff convergence following the augmented set estimator as in Chernozhukov et al. (2007). □