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

Khai X. Chiong — USC

Matthew Shum — Caltech

September 2016

**Machine Learning: What's in it for Economics?**

Becker-Friedman Institute, Univ. of Chicago

# Motivation

- Use machine learning ideas in *discrete choice models*
- Workhorse model of demand in economics and marketing.
- For applications in economics and marketing: hi-dim data
  - E-markets/platforms: Amazon, eBay, Google, Uber, Facebook, etc.
  - Large databases from traditional retailers (supermarket data)
- Many recent applications of these models face problem that consumers' choice sets are huge:
  - Where do Manhattan taxicab drivers wait for fares? (Buchholz 2016)
  - Legislators' choice of language (Gentzkow, Shapiro, Taddy 2016)
  - Restaurant choices in NYC (Davis, Dingel, Monras, Morales 2016)
  - Choice among *bundles* of products (eg. Fox and Bajari 2013)

# Specifically:

- **This paper:** address dimension-reduction of large choice set
  - (*not* large number of characteristics)[1]
- New application of *random projection* – tool from machine learning literature – to reduce dimensionality of choice set.
  - One of first uses in econometric modeling[2]
  - Use machine learning techniques in *nonlinear* econometric setting
- Semiparametric: Use *convex-analytic* properties of discrete-choice model (cyclic monotonicity) to derive inequalities for estimation[3]

[1] Chernozhukov, Hansen, Spindler 2015; Gillen, Montero, Moon, Shum 2015

[2] Ng (2016)

[3] Shi, Shum, Song 2015; Chiong, Galichon, Shum 2016; Melo, Pogorelskiy, Shum 2015

# Multinomial choice with large choice set

- Consider discrete choice model. The choice set is $j \in \{0, 1, 2, \ldots, d\}$ with $d$ being very large.
- Random utility (McFadden) model: choosing product $j$ yields utility

$$\underbrace{U_j}_{\text{utility index}} + \underbrace{\epsilon_j}_{\text{utility shock}} \quad \text{with } U_j = X_j'\beta$$

$X_j$ (dim $p \times 1$) denotes product characteristics (such as prices) and $\epsilon_j$ is utility shock (random across consumers).

- Highest utility option is chosen:

$$\text{choose } j \Leftrightarrow U_j + \epsilon_j \geq U_{j'} + \epsilon_{j'},\ j' \neq j$$

- $\beta$ (dim $p \times 1$) are parameters of interest.

# Discrete-choice model: assumptions and notation

- Notation:
  - $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)'$, $\quad \vec{U} = (U_1, \ldots, U_d)'$, $\quad \mathbf{X} \equiv (\vec{X}_1, \cdots, \vec{X}_d)'$
  - Market share (choice probability): for a given utility vector $\vec{U}$

$$s_j(\vec{U}) \equiv Pr(U_j + \epsilon_j \geq U_{j'} + \epsilon_{j'}, j' \neq j)$$

- Aggregate data: we observe data $\{\vec{s}^m, \mathbf{X}^m\}_{m=1}^M$ across markets $m$
- Assumptions:
  - Utility shocks are independent of regressors: $\vec{\epsilon} \perp \mathbf{X}$. No endogeneity.
  - Distribution of $\vec{\epsilon}$ is unspecified: *semiparametric.* Don't restrict correlation patterns among $\epsilon_j$, $\epsilon_{j'}$ (*may not be IIA*).
  - Normalize utility from $j = 0$ to zero.

# Convex analysis and discrete choice

- Since we don't specify distribution of $\vec{\epsilon}$, parametric DC models (MN logit, nested logit, etc.) aren't appropriate here.
- Instead, estimate using inequalities derived from convexity properties of discrete choice model.
- Namely, the expected maximal utility for decisionmaker (McFadden's "social surplus function")

$$\mathcal{G}(\vec{U}) = \mathbb{E}[\max_j(U_j + \epsilon_j)] \quad \text{is } \textit{convex} \text{ in } \vec{U}.$$

- Market shares at $\vec{U}$ correspond to (sub-)gradient Define of $\mathcal{G}$ at $\vec{U}$:[4]

$$\vec{s}(\vec{U}) \in \partial \mathcal{G}(\vec{U}).$$

We derive estimating inequalities from property of mkt shares:

[4] McFadden (1981). This is the (generalized) Daly-Zachary Theorem

# Estimating inequalities: Cyclic monotonicity

- Recall: (sub)-gradient of $\mathcal{G}(\vec{U})$ consists of mkt shares $\vec{s}(\vec{U})$.
- The (sub-)gradient of a (multivariate) convex function is **cyclic monotone**: for any cycle of markets $m = 1, 2, ..., L, L+1 = 1$

$$\sum_{m=1}^{L} (\vec{U}^{m+1} - \vec{U}^{m}) \cdot \vec{s}^{m} \leq 0 \quad \text{or} \quad \sum_{m} (\mathbf{X}^{m+1} - \mathbf{X}^{m})' \beta \cdot \vec{s}^{m} \leq 0.$$

  Inequalities do not involve $\epsilon$'s: estimate $\beta$ semiparametrically.[5]
- These inequalities valid even when some market shares=0
  - Empirically relevant (store-level scanner data)[6]
  - Consideration sets, rational inattention[7]

[5] Shi, Shum, and Song (2015); Melo, Pogorelskiy, Shum (2015)
[6] Gandhi, Lu, Shi (2013). We allow $\epsilon$ to have finite support.
[7] Matejka, McKay 2015

# Introducing random projection

- Problem: $\vec{U}^m$ and $\vec{s}^m$ are $d$ (very large) dimensional.
- Use **random projection** from $\mathbb{R}^d \to \mathbb{R}^k$, with $k << d$.
  - Consider: $d \times 1$-vector $\vec{y}$; Random matrix $\mathbf{R}$ ($k \times d$).
  - Projection is given by $\widetilde{y} = \frac{1}{\sqrt{k}}\mathbf{R}\vec{y}$, resulting in a $k \times 1$ vector.
  - Many candidates for $\mathbf{R}$; we consider *sparse random projection*[8]:

$$r_{i,j} \in \sqrt{\psi} \cdot \{+1, 0, -1\} \quad \text{with probs.} \left\{\frac{1}{2\psi}, 1 - \frac{1}{\psi}, \frac{1}{2\psi}\right\}$$

  - $\psi =$ "sparseness".
    - Eg. if $\psi = \sqrt{d}$, and $d = 5000$, use $< 2\%$ of data.

[8] Archiloptas 2003; Li, Hastie, Church 2006

# Properties of Random projection

- RP replaces high-dim vector $\vec{y}$ with random low-dim vector $\widetilde{y}$ with *same length* (on average): given $\vec{y}$, we have:

$$\mathbb{E}[\|\widetilde{y}\|^2] = \mathbb{E}[\|\mathbf{R}\vec{y}\|^2] = \|\vec{y}\|^2.$$

- Variance $V(\widetilde{y}) = O(1/k)$
- Use of random projection justified by the **Johnson-Lindenstrauss theorem:**

# Johnson-Lindenstrauss Theorem

- Consider projecting $d$-dim vectors $\{\vec{w}\}$ down to $k$-dim vectors $\{\widetilde{w}\}$;

There exists an $\mathbb{R}^d \to \mathbb{R}^k$ mapping which *preserves Euclidean distance among points*; ie. for all $m_1, m_2 \in \{1, 2, \ldots, M\}$ we have, for $0 < \delta < 1/2$ and $k = O(\log(M)/\delta^2)$

$$(1-\delta)\|\vec{w}^{m_1} - \vec{w}^{m_2}\|^2 \leq \|\widetilde{w}^{m_1} - \widetilde{w}^{m_2}\|^2 \leq (1+\delta)\|\vec{w}^{m_1} - \vec{w}^{m_2}\|^2.$$

The distance between the lower-dim vectors $(\widetilde{w}^{m_1}, \widetilde{w}^{m_2})$ lies within $\delta$-neighborhood of distance btw high-dim vectors $(\vec{w}^{m_1}, \vec{w}^{m_2})$.

- Proof is *probabilistic*: shows random projection achieves these bounds w/ positive prob.

# The RP Estimator

- Observed dataset: $\mathcal{D} \equiv \{\vec{s}^m, \mathbf{X}^m\}_{m=1}^M$
- Projected dataset: $\widetilde{\mathcal{D}_k} = \left\{\widetilde{s}^m = \mathbf{R}\vec{s}^m,\ \widetilde{\mathbf{X}}^m = (\mathbf{R}\vec{X}_1^m, \dots, \mathbf{R}\vec{X}_p^m)\right\}_{m=1}^M$. (Project $\mathbf{X}^m$ column-by-column.)
- Projected CM inequalities: for all cycles in $m \in \{1, 2, \dots, M\}$

$$\sum_m (\widetilde{U}^{m+1} - \widetilde{U}^m) \cdot \widetilde{s}^m = \sum_m (\widetilde{\mathbf{X}}^{m+1} - \widetilde{\mathbf{X}}^m)'\beta \cdot \widetilde{s}^m \leq 0$$

**The RP Estimator** $\widetilde{\beta}$ **minimizes the criterion function:**

$$Q(\beta, \widetilde{\mathcal{D}}) = \sum_{\text{all cycles}; L \geq 2} \left[ \sum_{m=1}^{L} \left( (\widetilde{\mathbf{X}}^{m+1} - \widetilde{\mathbf{X}}^m)'\beta \cdot \widetilde{s}^m \right) \right]_+^2$$

Convex in $\beta$ (convenient for optimization); may have multiple optima

# Properties of RP estimator

- Why does random projection work for our model?
- Exploit alternative representation of CM inequalities in terms of Euclidean distance between vectors:[9]

$$\sum_m \left( \| \widetilde{U}^m - \widetilde{s}^m \|^2 - \| \widetilde{U}^m - \widetilde{s}^{m-1} \|^2 \right) \leq 0$$

- By JL Theorem, RP preserves Euclidean distances between corresponding vectors in $\mathcal{D}$ and $\widetilde{\mathcal{D}}$.
- If CM inequalities satisfied in original dataset $\mathcal{D}$ should also be (approximately) satisfied in $\widetilde{\mathcal{D}}$.

[9] Villani 2003

# Properties of RP estimator (cont'd)

- RP estimator $\widetilde{\beta}$ is random due to
  1. randomness in **R**
  2. randomness in market shares $s_j^m = \frac{1}{N_m} \sum_i \mathbb{1}(y_{i,j} = 1)$
- For now, focus just on #1: highlight effect of RP
  - (Assume market shares deterministic; not faroff)
- Inference: open questions
  - We show uniform convergence of $Q(\beta, \widetilde{\mathcal{D}})$ to $Q(\beta, \mathcal{D})$ as $k$ grows. Building block for showing consistency of $\widetilde{\beta}$
  - For inference: little guidance from machine learning literature
  - In practice, assess performance of RP estimator across independent RP's
- Monte Carlo; Two applications 1.Scanner data 2.Mobile advertising

# Monte Carlo Simulations

- Designs: $d \in \{100, 500, 1000, 5000\}$; $k \in \{10, 100, 500\}$; $M = 30$
- In each design: fix data across replications, but redraw **R**. Report results across 100 independent RP's.
- Utility specification: $U_j = X_j^1 \beta_1 + X_j^2 \beta_2 + \epsilon_j$
  - Two regressors: $X^1 \sim N(1,1)$ and $X^2 \sim N(-1,1)$
  - Normalize $||\beta|| = 1$: set $\beta_1 = \cos\theta$, $\beta_2 = \sin\theta$ with true $\theta_0 = 0.75\pi = 2.3562$.
  - Random error structure: MA(2) serial correlation in errors across products (non MNL, non-exchangeable)
- Only using cycles of length 2 and 3 (similar results with longer cycles)

# Monte Carlo results

- Results are robust to different DGP's for RP
  - $\psi = 1 \Rightarrow$ Dense random projection matrix.
  - $\psi = \sqrt{d} \Rightarrow$ Sparse random projection matrix.
- In most cases, optimizing $Q(\beta, \widetilde{\mathcal{D}})$ yields a unique minimum.
- On average, estimates close to the true value, but there is dispersion across RP's.

# Monte Carlo results: Sparse random projection matrix

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

| Design | mean LB (s.d.) | mean UB (s.d.) |
|---|---|---|
| $d = 100, k = 10$ | 2.3073 (0.2785) | |
| $d = 500, k = 100$ | 2.2545 (0.2457) | 2.3473 (0.2415) |
| $d = 1000, k = 100$ | 2.3332 (0.2530) | 2.3398 (0.2574) |
| $d = 5000, k = 100$ | 2.3671 (0.3144) | |
| $d = 5000, k = 500$ | 2.3228 (0.3353) | 2.5335 (0.3119) |

Replicated 100 times using independently realized **sparse** random projection matrices. The true value of $\theta$ is 2.3562.

# Application I: Store and brand choice in scanner data

- Soft drink sales of Dominick's supermarkets ($\mathfrak{Rip}$) in Chicago
- Consumers choose both the type of soft drink and store of purchase.
- Leverage virtue of semiparametric approach.
  - Typically store/brand choice modelled as tiered discrete-choice model (i.e. nested logit).
  - Our approach: no need to specify tiering structure. Do consumers choose stores first and then brands, or vice versa?[10]
- $M = 15$ "markets" (two-week periods Oct96 - Apr97).
- Choose among 11 supermarkets (premium-tier and medium-tier).
- A choice is store/UPC combination: $d = 3060$ available choices.
- Reduce to $k = 300$ using random projection. Results from 100 independent RP's

[10] Hausman and McFadden (1984)

# Summary Statistics

| | Definition | Summary statistics |
|---|---|---|
| $s_{jt}$ | Fraction of units of store-upc $j$ sold during market (period) $t$ | Mean: 60.82, s.d: 188.37 |
| $price_{jt}$ | Average price of the store-upc $j$ during period $t$ | Mean: \$2.09, s.d: \$1.77 |
| $bonus_{jt}$ | Fraction of weeks in period $t$ for which store-upc $j$ was on promotion (eg. "buy-one-get-one-half-off") | Mean: 0.27, s.d: 0.58 |
| $holiday_t$ | Dummy variable for 11/14/96 to 12/25/96 (Thanksgiving, Christmas) | |
| $medium\_tier_j$ | Medium, non-premium stores.[a] | 2 out of 11 stores |
| $d$ | Number of store-upc | 3059 |
| $k$ | Dimension of RP | 300 |

Number of observations is 45885 = 3059 upcs × 15 markets (2-week periods).

# Empirical results

- Criterion function always uniquely minimized (but estimate does vary across different random projections)
- Purchase incidence decreasing in price, increasing for bonus, holiday
- Price coefficient negative
  - and lower on discounted items (*bonus*): more price sensitive towards discounted items
  - and lower during holiday season: more price sensitive during holidays
- No effect of store variables (*mediumtier*)

(Additional application: 2.Mobile advertising)

# Store/brand choice model estimates

Random projection estimates, dimensionality reduction from $d = 3059$ to $k = 300$.

| Specification | (C) | (D) |
|---|---|---|
| price | $-0.7729$ | $-0.4440$ |
| | $[-0.9429, -0.4966]$ | $[-0.6821, -0.2445]$ |
| bonus | $0.0461$ | $0.0336$ |
| | $[0.0054, 0.1372]$ | $[0.0008, 0.0733]$ |
| price $\times$ bonus | $-0.0904$ | $-0.0633$ |
| | $[-0.3164, 0.0521]$ | $[-0.1816, 0.0375]$ |
| holiday | $0.0661$ | $0.0238$ |
| | $[-0.0288, 0.1378]$ | $[-0.0111, 0.0765]$ |
| price $\times$ holiday | $-0.3609$ | $-0.1183$ |
| | $[-0.7048, -0.0139]$ | $[-0.2368, -0.0164]$ |
| price $\times$ medium_tier | | $0.4815$ |
| | | $[-0.6978, 0.8067]$ |
| | $d = 300$ Cycles of length 2 & 3 | |

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.

# Remarks

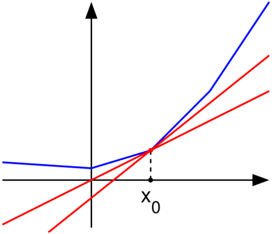
- For RP estimation, all that is needed is projected dataset $\widetilde{\mathcal{D}}$. Never need original dataset. Beneficial if **privacy** is a concern.
- Other approaches to large choice sets
  1. Multinomial logit with "sampled" choice sets.[11]
  2. Maximum score semiparametric approach.[12] Use only subset of inequalities implied by DC model.
     - ★ Estimation based on *rank-order property* (pairwise comparisons among options)
     - ★ In binary choice case: CM and ROP coincide.
     - ★ For multinomial choice: CM and ROP assumptions non-nested and non-comparable. Details
  3. Moment inequalities.[13] Omnibus method

[11] McFadden (1978); Ben-Akiva, McFadden, Train (1987)
[12] Fox (2007); Fox and Bajari (2013)
[13] Pakes, Porter, Ho, Ishii (2015)

# Conclusions

- Multinomial choice problem with huge choice sets
- New application of machine learning tool (random projection) for dimension reduction in these models.
- Derive semiparametric estimator from cyclic monotonicity inequalities.
- Procedure shows promise in simulations and in real-data application.
- Random projection may be fruitfully applied in other econometric settings
- Thank you!

# Convex analysis: subgradient/subdifferential/subderivative




- Generalization of derivative/gradient for nondifferentiable functions
- The *subgradient of* $\mathcal{G}$ *at* $p$ are vectors $u$ s.t.

$$\mathcal{G}(p) + u \cdot (p' - p) \leq \mathcal{G}(p'), \qquad \text{for all } p' \in \mathbf{dom}\,\mathcal{G}.$$

- Dual relationship between $u$ and $p$:
  - $\partial\mathcal{G}(p) = \operatorname{argmax}_{u \in \mathbb{R}^{|\mathcal{Y}|}} \{p \cdot u - \mathcal{G}^*(u)\}$,

    where $\mathcal{G}^*(u) = \max_{p \in \Delta^{|\mathcal{Y}|}} \{u \cdot p - \mathcal{G}(p)\}$. (Lemma)

# Remark: Other approaches to large choice sets

1. Maximum score semiparametric approach.[14] Use only subset of inequalities implied by DC model.
   - Estimation based on *rank-order property*: for all choices $j \neq j'$, pairwise comparisons characterize optimal choice:

   $$s_j > s_{j'} \leftrightarrow \mathbf{X}_j'\beta > \mathbf{X}_{j'}'\beta.$$

   - In binary choice case: CM and ROP coincide.
   - For multinomial choice: ROP implied by *exchangebility of* $F_{\epsilon|\mathbf{X}}$ (restrictions on correlation among $\epsilon_{j'}, \epsilon_j$, etc.)
   - In contrast, we assume independence $\epsilon \perp \mathbf{X}$ but leave correlation structure among $\vec{\epsilon}$ free. Non-nested and non-comparable.

Back

[14] Fox (2007); Fox and Bajari (2013)

# Application II: Choosing advertisers in mobile app markets

Back

- Model matching in online app market (joint with Richard Chen)
- Sellers: publishers sell "impressions" (users of online apps)
- Buyers: advertisers who vie to show mobile ad to user. Advertisers bid "cost-per-install" (CPI); only pay when user installs app.
- Data from major mobile advertising intermediary: chooses the optimal ads from one side to show to users on the other side.
- Intermediary wants to constantly evaluate whether optimality is achieved. *Optimality means choosing advertisers bringing high expected revenue*. Are these advertisers being chosen?
- However, difficult to do under CPI mechanism.
  - CPI payment may benefit advertisers (offering them "free exposure") but hurts publishers[15]

[15] Hu, Shin, Tang 2016

# Application II: Data Back

- Data from a major mobile app advertising intermediary
- Estimate model of probability that an advertiser gets chosen in US-iOS market.
- $>$7700 advertisers. Reduce to 1000.
- Advertiser covariates:
  - Lagged revenue (measure of expected revenue)
  - Lagged conversion probability (whether ad viewers install app)
  - Genre: gambling
  - Self-produced ad
  - Whether app is available in Chinese language

# Application II: Results Back

| Specification | (A) | (B) | (C) | (D) |
|---|---|---|---|---|
| Revenues | 0.823 (0.147)<br>[0.722, 0.937] | 0.521 (0.072)<br>[0.494, 0.563] | 0.663 (0.263)<br>[0.711, 0.720] | 0.657 (0.152)<br>[0.625, 0.748] |
| ConvProb | 0.069 (0.547)<br>[-0.445,0.577] | 0.037 (0.183)<br>[-0.076,0.161] | 0.006 (0.035)<br>[-0.013,0.033] | 0.025 (0.188)<br>[-0.112,0.168] |
| Rev $\times$ Gamble | | -0.809 (0.187)<br>[-0.856,-0.813] | -0.200 (0.098)<br>[-0.232,-0.185] | -0.192 (0.429)<br>[-0.500,0.029] |
| Rev $\times$ Client | | | -0.604 (0.278)<br>[-0.673,-0.652] | |
| Rev $\times$ Chinese | | | | -0.489 (0.228)<br>[-0.649,-0.409] |
| | Dimension reduction: $k = 1000$<br>Sparsity: $s = 3$<br>Cycles of length 2 and 3 | | | |

Table: Random projection estimates, $d = 7660$, $k = 1000$.

First row in each entry present the mean (std dev) 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.

# Application II results and discussion Back

- Robust results: expected revenues has strong positive effect, but conversion probability has basically zero effect. Once we control for revenues, it appears that conversion probability has no impact.
- Gamble, client and Chinese all mitigate the effect of revenues. Revenues appear less important for an advertiser when it is a gambling app, the creative is self-produced, or if app is available in Chinese.
- Are "right" advertisers being chosen?
  - Yes, to some extent: advertisers offering higher expected revenue are chosen with higher probability.
  - Partially reversed for gambling apps, self-produced ads– sub-optimal?