2022 Federal Trade Commission Microeconomics Conference: Diversion and the Use of Second-Choice Data

Julie Holland Mortimer (University of Virginia & NBER)

November 3rd, 2022

Review of Diversion Ratios

Diversion Ratios

The diversion ratio is one of the best ways we have to measure competition between products.

- Raise the price of product j and count the number of consumers who leave
- The diversion ratio $D_{j\to k}$ is the fraction of leavers who switch to the substitute k.
- A higher value of $D_{j\to k}$ indicates closer substitutes.
- Useful because it arises in the multi-product Bertrand FOC:

$$\underbrace{p_j\left(1 + 1/\epsilon_{jj}(\mathbf{p})\right)}_{\text{Marginal Revenue}} = c_j + \sum_{k \in \mathcal{J}_f \setminus j} (p_k - c_k) \cdot D_{j\to k}(\mathbf{p}).$$

- $D_{j\to k} \equiv \frac{\partial q_k}{\partial p_j} / \left| \frac{\partial q_j}{\partial p_j} \right|$.
- Can also write as $D_{j\to k} \equiv \frac{\epsilon_{kj}}{|\epsilon_{jj}|} \cdot \frac{q_j}{q_k}$

General Advantages of Diversion

- Diversion allows for unit-free comparisons (shares sum to one).
  - While own-elasticities are unit-free, this is not true of cross-elasticities.
  - Is $\epsilon_{jk} = .01$ or $\epsilon_{jk} = .03$ a better substitute? We can't tell.
    - Need $\epsilon_{jk} \cdot s_k$ to know.
    - But $\epsilon_{jk} \cdot \frac{s_k}{p_j} = D_{j \to k}$.
    - The fraction of switchers choosing k allows comparisons.
  - If tempted to report cross elasticities, consider reporting diversion ratios instead.
- Data on diversion can provide helpful variation for demand estimation.
  - Petrin (2002), MicroBLP (2004), Grieco, Murry, Yurukoglu (2022)
- Diversion can be a helpful complement to merger simulation.

Advantages of Diversion over Concentration (Farrell and Shapiro, 2010)

Diversion vs. concentration:

- Most goods and services are differentiated.
- Merger policy should aim to measure the substitutability of the differentiated offerings of competing firms.
- Concentration measures typically struggle to do this:
  - not all firms "in the market" produce products that are equally good substitutes
  - some firms "outside the market" may produce products that compete.
- If merging parties know they compete more closely than market-share analysis would predict, we'll have under-enforcement.

Diversion as a Treatment Effect (Conlon Mortimer RJE 2021)

Diversion Ratio = fraction of consumers who switch from purchasing a product j to purchasing a substitute k *(following an increase in the price of j)*

Treatment not purchasing product j

Outcome fraction of consumers who switch from $j \rightarrow k$.

Compliers consumers who would have purchased at z_j but do not purchase at z_j'.

This admits a Wald estimator:

$$D_{j \rightarrow k}(x) = \frac{\mathbb{E}[q_k | Z = z_j'] - \mathbb{E}[q_k | Z = z_j]}{\mathbb{E}[q_j | Z = z_j] - \mathbb{E}[q_j | Z = z_j']}$$

A LATE Theorem (Conlon Mortimer RJE 2021)

We also showed that most discrete-choice models yield the following representation:

$$D_{j\to k}^{z_j \to z_j'}(x) = \int_{z_j}^{z_j'} D_{j\to k,i}(x)\, w_i(z_j, z_j', x)\, dF_i \text{ with } w_i(z_j, z_j', x) = \frac{s_{ij}(z_j, x) - s_{ij}(z_j', x)}{s_j(z_j, x) - s_j(z_j', x)}$$

- Different interventions $z_j \to z_j'$ (prices, quality, characteristics, assortment) give different weights $w_i(z_j, z_j', x)$ and thus different local average diversion ratios.
- Individual Diversion Ratios $D_{j\to k,i}(x)$ don't vary with the intervention (determined only by how i ranks 2nd and 3rd choices).
- That paper establishes the decomposition above and derives some properties.

Properties of Diversion Ratios (Conlon Mortimer RJE 2021)

$$\int_{z_j}^{z_j'} D_{j\to k,i}(x) w_{ij}(z_j, z_j', x)\partial F_i = \int_{z_j}^{z_j'} \frac{s_{ik}(x)}{1 - s_{ij}(x)} w_{ij}(z_j, z_j', x)\partial F_i = \int s_{ik}(x)\, \widetilde{w}_{ij}(z_j, z_j', x)\partial F_i$$

where s_{ij} is probability that i chooses j or $Pr(u_{ij} > u_{ik})$ for all $k \in \mathcal{J}$ and $k \neq j$

- For any (mixed) logit $D_{j\to k,i}(x) = \frac{s_{ik}}{1-s_{ij}}$
- For plain logit $D_{j\to k,i} = \frac{s_k}{1-s_j}$ for all i
  - imposes constant diversion
  - weights don't matter

Weights (Conlon Mortimer RJE 2021)

Weights on Treatment Effects parameters for RC logit:

	$\widetilde{w}_{ij}(z_j, z_j', x) \propto$
second-choice data	$\frac{s_{ij}(x)}{1-s_{ij}(x)}$
price change $\frac{\partial}{\partial p_j}$	$s_{ij}(x) \cdot \lvert\alpha_i\rvert$
characteristic change $\frac{\partial}{\partial x_j}$	$s_{ij}(x) \cdot \lvert\beta_i\rvert$
small quality change $\frac{\partial}{\partial \xi_j}$	$s_{ij}(x)$
finite price change $w_i(p_j, p_j', x)$	$\frac{\lvert s_{ij}(p_j', x) - s_{ij}(p_j, x)\rvert}{1-s_{ij}(x)}$
finite quality change $w_i(\xi_j, \xi_j', x)$	$\frac{\lvert s_{ij}(\xi_j', x) - s_{ij}(\xi_j, x)\rvert}{1-s_{ij}(x)}$

Price interventions put more weight on the most price-sensitive types,
Quality interventions put more weight on the most quality-sensitive types, etc.

Motivating the Use of Second-Choice Data

Estimating Preferences and Substitution Patterns from Second-Choice Data

Joint work with Christopher T. Conlon (NYU) and Paul Sarkis (Boston College)

There are many cases where we observe second-choice data: (the probability that i chooses k as their second choice conditional on choosing j as their first choice):

- Rank-ordered lists (market design, school choice)
- Customer Surveys: (If you didn't buy a Camry what would you buy?)
- Conjoint analyses in Marketing
- A/B tests showing different search results to different customers.

Research Question (Conlon, Mortimer, Sarkis)

We consider a problem where we observe some aggregate shares $\mathcal{S} = [\mathcal{S}_1, \ldots, \mathcal{S}_J]$ or sales $\mathcal{Q}_j$, and some elements $(j,k) \in \text{OBS}$ of $\mathcal{D}^T$ a matrix of (second-choice) diversion ratios.

$$\mathcal{D}^T = \begin{array}{c} \begin{array}{ccccc} \text{VZ} & \text{ATT} & \text{TMo} & \text{S} & \text{Other} \end{array} \\ \left(\begin{array}{ccccc} 0 & ? & 0.30 & 0.30 & ? \\ ? & 0 & 0.45 & 0.15 & 0 \\ ? & ? & 0 & 0.45 & ? \\ ? & ? & 0.20 & 0 & ? \\ ? & ? & 0.05 & 0.10 & 0 \end{array} \right) \end{array} \begin{array}{c} \text{VZ} \\ \text{ATT} \\ \text{TMo} \\ \text{S} \\ \text{Other} \end{array}, \quad \begin{bmatrix} 0.35 \\ 0.30 \\ 0.20 \\ 0.15 \\ 0.05 \end{bmatrix} = \mathcal{S}$$

Can we fill in the missing elements?

How do we fill in missing elements?

Typical Approach: estimate a parametric model.

- Multi-product demand with unrestricted matrices of $(J+1)^2$ cross-elasticities (such as AIDS) is often hopeless with large J. Unrestricted diversion likely equally hopeless.
- Plain logit places strong restrictions: $D_{j \to k} = \frac{s_k}{1-s_j}$.
- Nested logit $D_{j \to k} = \frac{s_{k|g}}{Z(\sigma, s_g) - s_{j|g}}$ (same nest) where σ is nesting parameter.

How do we fill in missing elements?

Mixed Logit: Explain substitution patterns using observed characteristics

- Typically assume independent normal RC
- Two products with similar x_1 and high substitution → larger σ_1.
- Two products with similar x_2 and low substitution → smaller σ_2.

McFadden and Train (2000) show a mixed logit $u_{ij} = \beta_i x_j + \varepsilon_{ij}$ is fully flexible

1. This depends on $f(\beta_i)$ heterogeneity being nonparametric
2. And a sufficient set of characteristics X to explain $\mathcal{D}$

Much work on (1), less attention on (2).

How do we fill in missing elements?

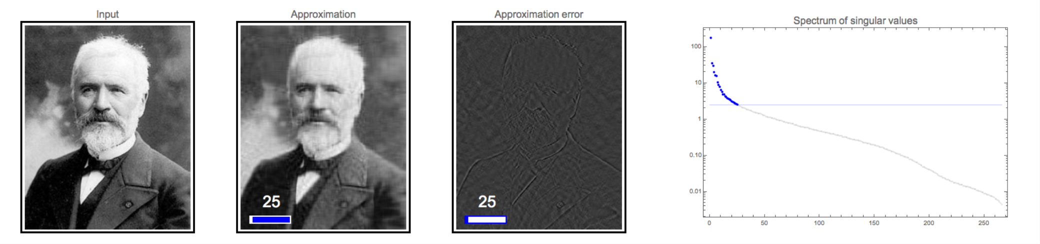
Our paper: Consider a low-rank approximation to $\mathcal{D}$

- Limit the rank of $\mathcal{D}$ directly in product space instead of controlling complexity with product characteristics and parametric restrictions on random coefficients.
- Allow for sparsity in individual shares and substitution patterns, with possibility of generating extreme patterns for top substitutes if necessary.

Works well in other domains (CS for image recovery/compression), and we show it has a sensible economic interpretation.

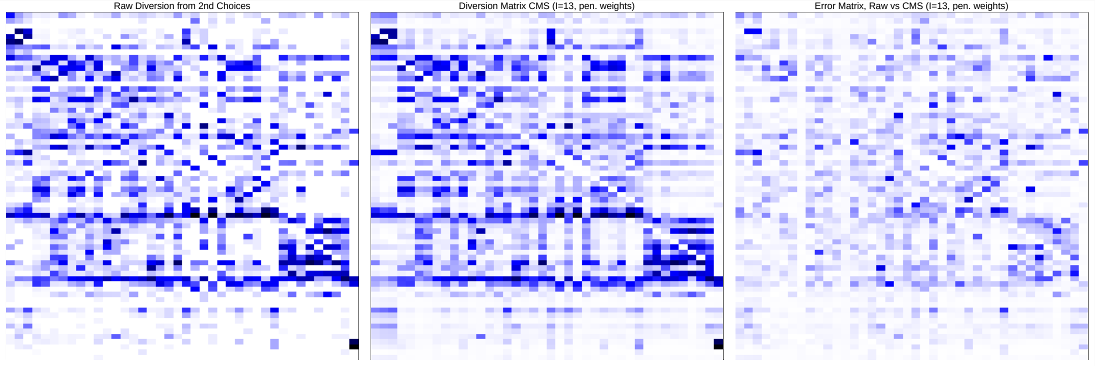
Low Rank Approximations: Image Compression

Image of Camille Jordan (1838-1922)



$$A \approx U_{266\times 25} \cdot \Sigma_{25\times 25} \cdot V_{25\times 266}$$

Completing the Matrix: $\mathcal{D}$ for Autos



When might we want to do this?

- We have access to aggregate market shares and some (but not all) second-choice data (microBLP (2004); Grieco, Murry, Yurukoglou (2022)).
- We are interested in estimating substitution patterns across all sets of products but have data on only a subset
  - shares of largest cellular phone providers, and number porting or switching data for merging parties only.
  - survey data on "If this Tesco were to close where would you shop" (as UK CMA asks).
  - win-loss data from merging parties only (Qiu, Sawada, Sheu (2022))
- We lack sufficient variation in prices, other covariates, to estimate demand system.
- Product characteristics do not accurately capture substitution across products.

Setup and Model

Assumptions

- Consumers make discrete choices from set $\mathcal{J}$
- Utility is given by semi-parametric logit

$$u_{ij} = V_{ij} + \varepsilon_{ij}$$
$$s_{ij} = Pr(u_{ij} > u_{ik}) \text{ for all } k \in \mathcal{J}, k \neq j.$$

- ε_{ij} is Type I extreme value.
- Goal: estimate $f(V_{ij})$.
- Strategy: Approximate with finite mixture with weights π_i.

Linear Algebra Notation

- Individual i's share for each choice given by $\mathbf{s_i} = [s_{i0}, s_{i1}, \ldots, s_{iJ}]$.
- Aggregate shares by $\sum_{i=1}^{I} \pi_i \cdot \mathbf{s_i} = \mathbf{s}$.
- The matrix of individual diversion ratios is given by $\mathbf{D}_i = \mathbf{s_i} \cdot \left[\frac{1}{(1-\mathbf{s_i})}\right]^T$.

We write the $(J+1) \times (J+1)$ matrix of second-choice diversion as:

$$D_{j\to k} = \sum_{i=1}^{I} \pi_i \cdot D_{j\to k,i} \cdot w_i = \sum_{i=1}^{I} \pi_i \cdot \frac{s_{ik}}{1-s_{ij}} \cdot \frac{s_{ij}}{s_j}$$

$$\mathbf{D} = \left(\sum_{i=1}^{I} \pi_i \cdot \mathbf{s_i} \cdot \left[\frac{1}{(1-\mathbf{s_i})}\right]^T \cdot \operatorname{diag}(\mathbf{s_i})\right) \cdot \operatorname{diag}(\mathbf{s})^{-1}$$

$$= \left(\sum_{i=1}^{I} \pi_i \cdot \mathbf{s_i} \cdot \left[\frac{\mathbf{s_i}}{(1-\mathbf{s_i})}\right]^T\right) \cdot \operatorname{diag}(\mathbf{s})^{-1}$$

Notation continued

Under relatively general conditions, second-choice diversion can be written as:

$$\text{diag}(\mathbf{s}) \cdot \mathbf{D} = \sum_{i=1}^{I} \pi_i \cdot \begin{bmatrix} | \\ \mathbf{s_i} \\ | \end{bmatrix} \cdot \begin{bmatrix} - & \frac{\mathbf{s_i}}{1-\mathbf{s_i}} & - \end{bmatrix}$$

- Each individual diversion ratio is of rank one since it is the outer product of $\mathbf{s_i}$ with itself (and some diagonal "weights").
- The (unrestricted) matrix of diversion ratios $\mathbf{D}$ is $(J+1) \times (J+1)$.
- Logit restricts $\mathbf{D}$ to be of rank one. Nested logit of rank $\leq G$ (the number of non-singleton nests). Mixed logit to $\text{rank}(\mathbf{D}) \leq I$ (but bound is likely uninformative).

Setting

- Assume that we observe aggregate market shares $\mathcal{S}_j$ and some subset of the diversion matrix $\mathcal{D}_{j \to k}$ for $(j,k) \in \text{OBS}$.
- Goal: Can we obtain an estimate for the remainder of the matrix $\mathcal{D}$?
    - Related to CS literature on matrix completion methods.
    - Useful tip from linear algebra: nuclear norm: $||A||_* = \sum_i \sigma_i(A)$ where $\sigma_i(A)$ are singular values. This works like a continuous approximation to rank.
    - We don't need to do nuclear norm penalization since discrete choice provides enough structure.
- Low-rank approximation is consistent with utility maximization under discrete choice.
    - Theoretical interpretation as indirect utilities, not just mech. rank reduction (ie: PCA).

Our Semiparametric Problem

$$\min_{s_{ij},\pi_i} \sum_{(j,k)\in\text{OBS}} \widetilde{c}_j \left(\mathcal{D}_{j\to k} - D_{j\to k}\right)^2 + \sum_j c_j \left(\mathcal{S}_j - \sum_i \pi_i \cdot s_{ij}\right)^2 + \lambda \left\|\pi_i\right\|^2$$

$$\text{subject to} \quad D_{j\to k} = \sum_{i=1}^{I} \pi_i \cdot \frac{s_{ik}}{1 - s_{ij}} \cdot \frac{s_{ij}}{s_j}$$

$$0 \le s_{ij}, \pi_i, s_j, D_{j\to k} \le 1, \quad \sum_{i=1}^{I} \pi_i = 1, \quad \sum_j s_{ij} = 1$$

- Use cross validation to select # of types I.
- With $\lambda > 0$ we penalize HHI of w_i and becomes elastic net
- Weights $\widetilde{c}_j$ and c_j are proportional to $\ln q_j$

Discussion

- Goal: a good predictive model for unobserved elements of $\mathcal{D}$.
- We are worried about overfitting so we use cross validation (withholding columns of $\mathcal{D}$) to select number of types I.
    - Otherwise we would always prefer the more complicated model
    - Compare models based on out-of-sample fit (RMSE, MAD).
- Model may or may not be sparse $s_{ij} = 0$ for some (i, j)
    - Could be that consumer i doesn't consider j.
    - Or consequence that $s_{ij} \geq 0$ and $\sum_j s_{ij} = 1$ amounts to an L_1 penalty $\sum_j |s_{ij}| \leq 1$
- Model is a semiparametric logit for $V_{ij} \in \mathbb{R}$ (don't rule out $V_{ij} \to \pm\infty$):

$$u_{ij} = V_{ij} + \varepsilon_{ij}, \quad s_{ij} = \frac{e^{V_{ij}}}{1 + \sum_k e^{V_{ik}}}$$

Comparison: Fox, Kim, Ryan, Bajari (QE 2011)

$$\min_{\pi_i \geq 0} \sum_j \left(\mathcal{S}_j - \sum_i \pi_i \cdot \hat{s}_{ij}(\widehat{\beta}_i) \right)^2 \quad \text{subject to} \quad \sum_i \pi_i = 1$$

$$\hat{s}_{ij}(\widehat{\beta}_i) = \frac{e^{\widehat{\beta}_i x_j}}{1 + \sum_{j'} e^{\widehat{\beta}_i x_j'}}$$

- Draw $\beta_i \sim G(\beta_i)$ from a prior distribution.
- Solved in characteristic space with a semi-parametric form for $F(\beta_i)$.
- Often produces very sparse models $\pi_i = 0$ (for all but 50 of 1000 simulated consumers).

Comparison: Raval et al. (2017, 2020)

- Cut data into bins (zip, income, age, gender)
- Observe shares (hospital demand) within each bin $s_{g(i),j}$
- A separate plain logit for each bin with only ξ_j as the common parameter.
- Use second choices from hospital closures (natural disasters) to compare models.

$$s_{g(i),j} = \frac{e^{\beta_g x_j + \xi_j}}{1 + \sum_{j'} e^{\beta_g x_{j'} + \xi_{j'}}}, \quad D_{j \to k,i} = \frac{s_{g(i),k}}{1 - s_{g(i),j}}$$

Comparison: Latent Class Logit (Greene and Hensher 2003)

Most similar to what we're doing.

- Estimate separate β_i for each class.
- Estimate proportion of each class π_i.
- Estimating finite mixtures is tricky and usually requires EM.

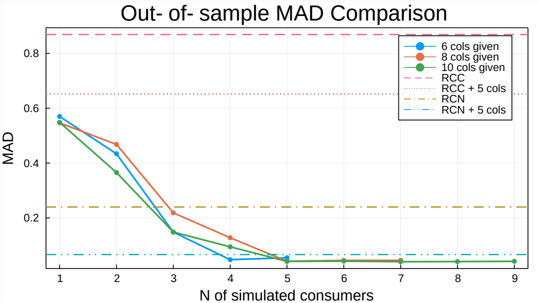
$$s_k(\pi, \beta) = \sum_{i=1}^{I} \pi_i \cdot \left(\frac{e^{\beta_i x_{ij} + \xi_j}}{1 + \sum_k e^{\beta_i x_{ik} + \xi_k}} \right)$$

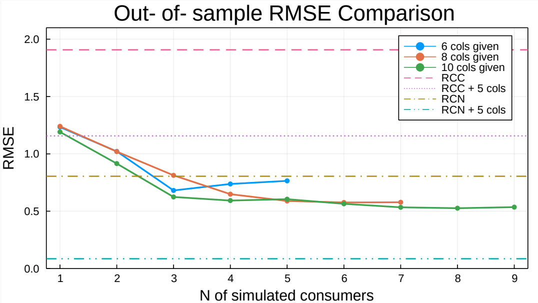
Monte Carlo

Generating Data

- Fit (i) nested logit, (ii) RC logit to data on vending machines from Conlon and Mortimer (JPE, 2021).
- Generate fake sales and diversion from those parameter estimates.
  - $J = 45$ products; $T = 250$ markets; with 30 randomly selected products in each. Market size $M = 1000$ per market. Nesting parameter is $\rho = 0.25$.
  - Categories: Salty Snacks, Chocolate, Non-Chocolate Candy, Cookies, Pastry, Other.
- Estimate a variety of misspecified parametric models: RC on nest dummies, RC on characteristics (Salt, Sugar, Nut Content), and our semiparametric estimator.
  - Include $m \ll J$ columns of $\mathcal{D}_{j\to k}$ as extra moments.
- Compare out-of-sample predicted Diversion Ratios.
  - MAD: Median $\left(|\mathcal{D}_{j\to k} - \hat{D}_{j\to k}|\right)$ for $(j,k) \in \{\text{Validation}\}$.
  - RMSE: $\sqrt{\frac{1}{n}\sum_{(j,k)\in\{\text{Validation}\}} |\mathcal{D}_{j\to k} - \hat{D}_{j\to k}|^2}$

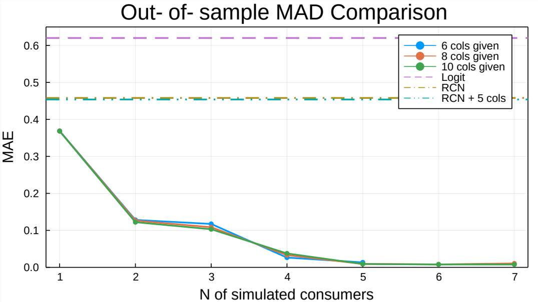
Monte Carlo: DGP is Nested Logit

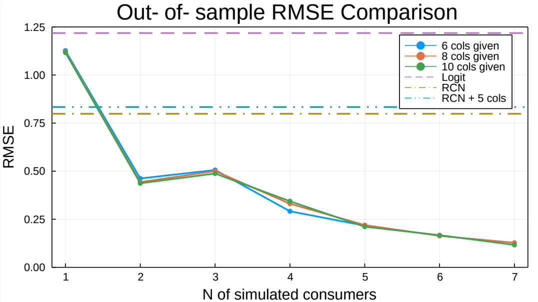




- RCC is mis-specified
- Diversion Moments improve efficiency of RCN
- $I \geq 4$ does a pretty good job.

Monte Carlo: DGP is RC on chars





- RCN is mis-specified
- $I \geq 4$ does a pretty good job.

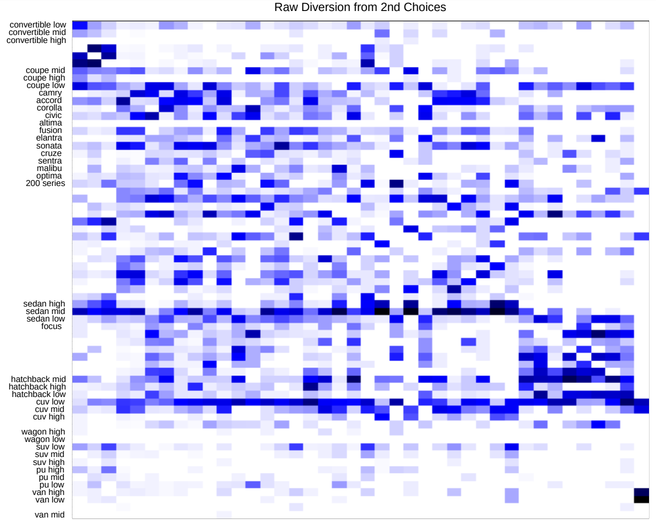
Application to Autos Data

Description of Autos Data

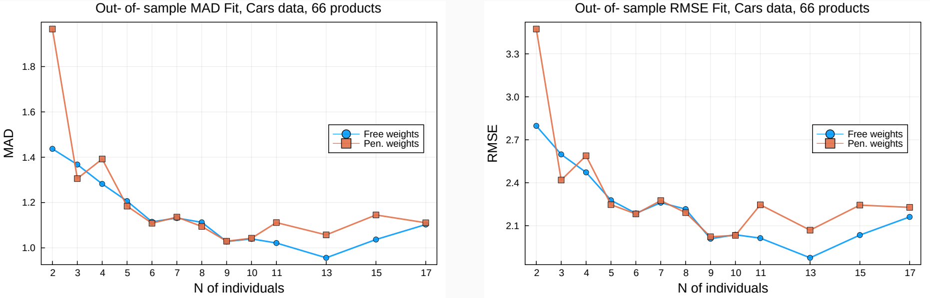
- Subset of data from Grieco, Murry and Yurukoglu (2022).
- Focus on one year of sales from 2015
  - Aggregate sales observed at the model-year level from Ward's Automotive.
  - Second choices from MaritzCX survey (53,328 purchases)
  - Construct $J = 181$ products by consolidating all models below 15,000 annual sales.
    - Consolidated products are: Car/Truck by Low/Mid/High prices (6 products)
- Same Goal: Predict unobserved second-choice data without characteristics.

MaritzCX Survey data (173 Cars and Trucks)

MaritzCX Survey data (66 Cars)



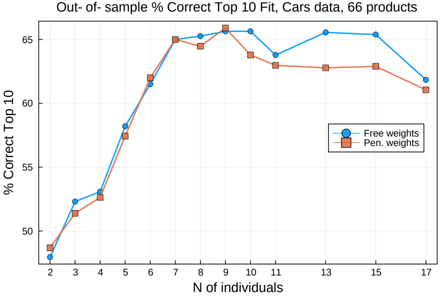
Cross Validation: Model Selection

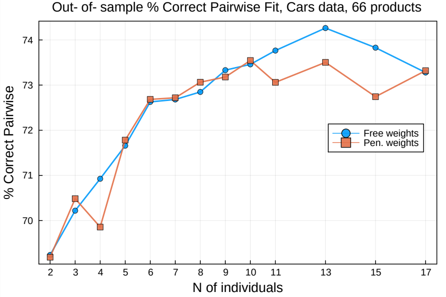


Dots are cross-validated means.

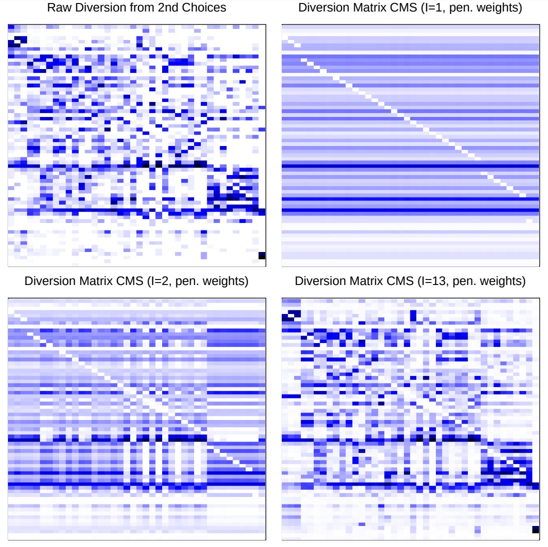
Seems to select $I = 13$ (bias-variance tradeoff).

Cross Validation: Model Selection





MaritzCX Survey data (66 Cars)

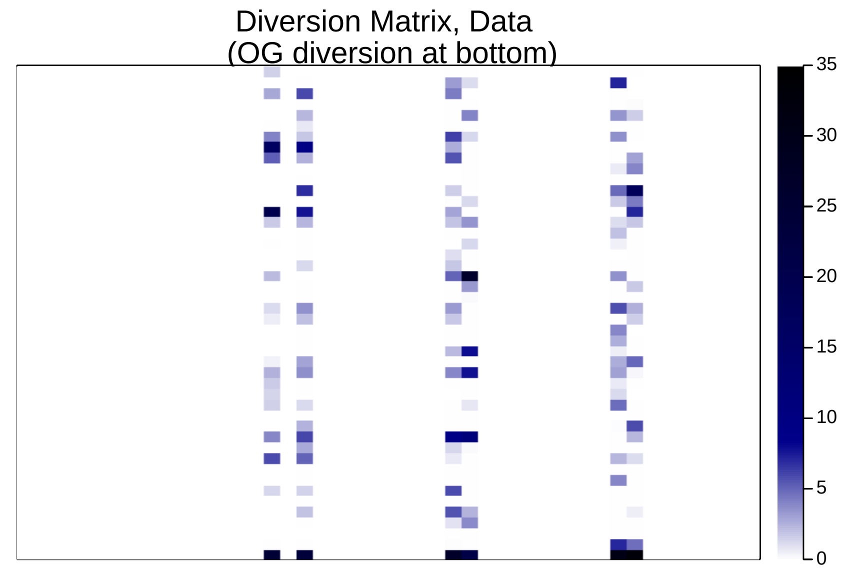


Application to Vending Data

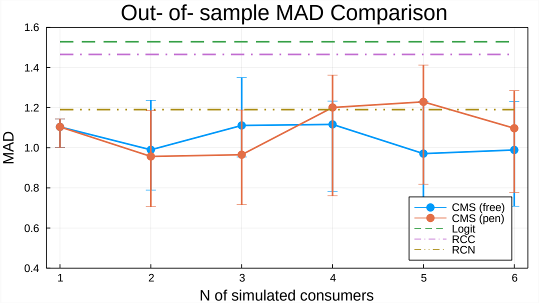
Description of Vending Data

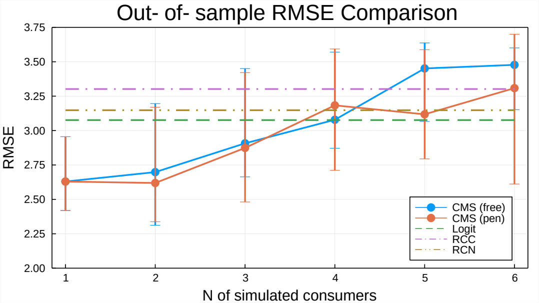
- Same data as Conlon and Mortimer (JPE, 2021).
- 66 Vending Machines in white-collar office buildings in downtown Chicago
- About 35-40 snack products in each building
- 6 exogenous product removals (2.5-3.5 weeks long each)
  - Snickers, M&M Peanut, Doritos Nacho, Cheetos, Animal Crackers, Famous Amos

Diversion Observed for 6 Products:



Cross Validation: Model Selection



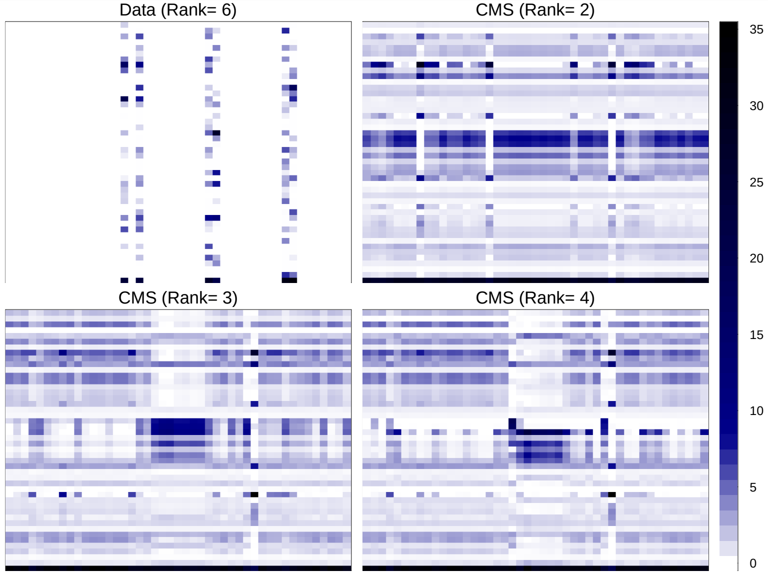


Out-of-sample fit (mostly) beats in-sample fit of parametric models.

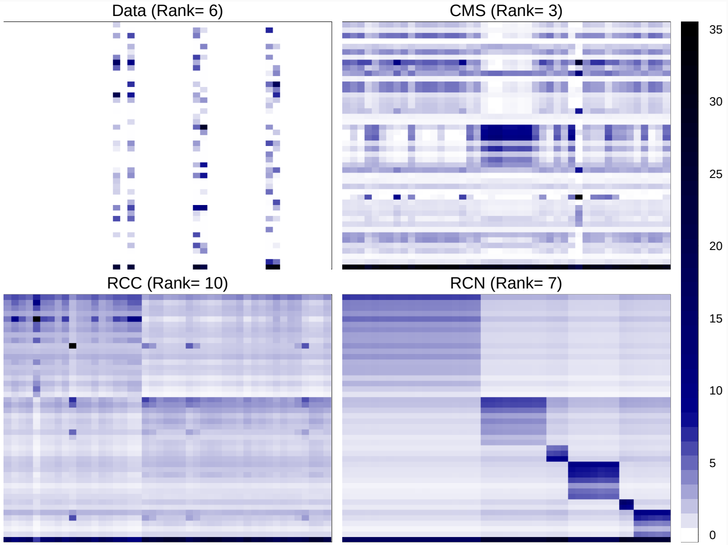
Error bars are across all holdout experiments/ Dots are cross-validated means.

Seems to select $I = 2$ or $I = 3$ (bias-variance tradeoff).

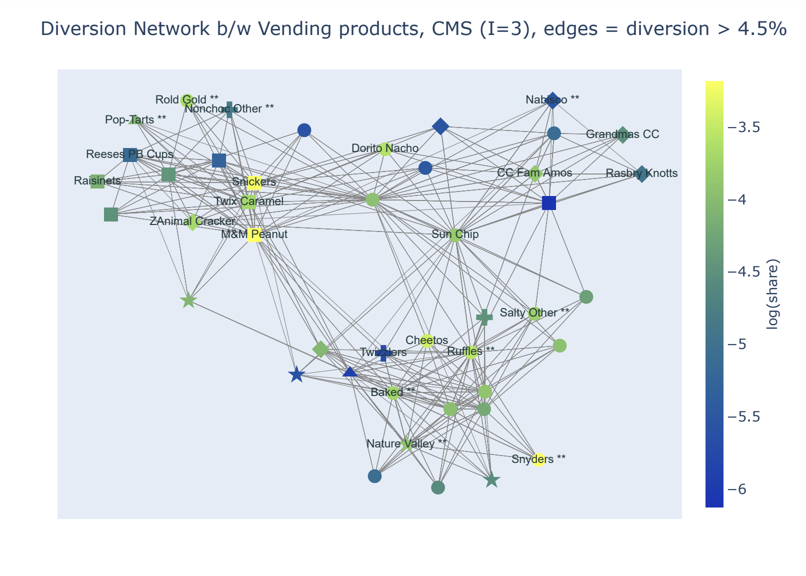
Diversion Matrix: Estimates Comparison $\lambda = 0$



Diversion Matrix: Estimates Comparison $\lambda = 0$



Network Structure of Vending Products: Semiparametric $I = 3$



Extensions and Conclusion

Extensions

- What about (exogenous) price or quality changes?
  Expression for $D_{j \to k}$ changes slightly.
- Want to add covariates or endogenous prices?
  Straightforward to run an IV regression:

$$\log s_{ij} - \log s_{i0} = x_j \beta_i + \xi_j$$

  Test how much we lose using only a basis in $f(x_1, x_2)$.
- Optimal Experimentation: Which product is most informative about $\mathcal{D}$?
  - $\mathcal{D}$ looks like a transition matrix with a network structure
  - Relates to measures of centrality / eigenvalues.
  - Cross elasticities are not a well-behaved network.

Conclusion

- Allowing for flexible unobserved types can give more accurate substitution patterns
  - Particularly true in capturing closeness of best substitutes not captured by product characteristics (e.g. Snickers and Peanut M&M's vs Snickers and Milky Way)
- Using observable substitution patterns (experiments or surveys) and "completing" the $(J+1) \times (J+1)$ matrix with a low-rank approximation looks promising.
- How much information on second choices is "enough"?
- Which products are important for completing substitution patterns?