Counterfactual Inference for Consumer Choice With Many Products

Susan Athey, Stanford University

Counterfactual Inference for Consumer Choice Across Many Product Categories (Susan Athey, David Blei, Rob Donnelly, Francisco Ruiz, in progress)
Consumer Choice
SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements (Francisco Ruiz, Susan Athey, David Blei, 2017)
Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data (Susan Athey, David Blei, Rob Donnelly, Francisco Ruiz, Tobias Schmidt, AEA Papers and Proceedings, 2018)

Asking and Answering Questions Using Panel Data with Consumer Choices Over Many Products

- Example Data Sets
  - List of Websites/Apps/News Articles Viewed
  - Device Id and Lat/Long of Locations Visited
  - Consumer Credit Card, Bank Transactions w/ Description
- Example Questions
  - How do users change their consumption of news when Google News shuts down, during election, when reading from Facebook, etc.?
  - How do physical movements of consumers change when they lose a job, when a store opens or closes?
  - How do consumers or suppliers in “gig economy” change spending patterns, travel, etc. as a result of the entry/increase in supply/changes in wages of Uber, Rover, etc.?
  - What products make good “loss leaders”? Interact w/ other products?
- Common features
  - Limited “structured” data about objects consumed
  - Consumers consumer wide variety of products, many rarely

Towards A Large Scale Model of Consumer Choice

One product at the time misses many aspects of consumer choice. E.g. for supermarkets:

- Store v. store competition happens at the level of a shopping trip, not an item
- Stores desire to understand profile of most valuable consumers, and attract them to the store
- Stores make decisions about products to stock and promote and how to price in order to attract different types of consumers
- Store organization can be made more or less convenient for different collections of products
- Bundling, loss leader strategies

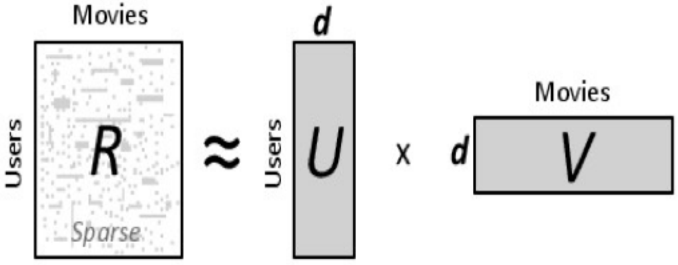
Economics/Marketing Literature Approaches to Demand

- Build a functional form model of user utility; estimate preferences based on user choice behavior
  - One product at the time (e.g. yoghurt)
  - Single product-specific latent variable (e.g. product quality) that may be correlated with price
  - Latent consumer preferences for observable product characteristics
  - Small number of papers: unobserved latent product characteristics (Goettler and Shachar, 2001; Athey-Imbens 2007; Nair, Misra, et al 2013)
- To identify effects of price
  - Instrumental variables approaches in cross-sectional data
  - Variety of approaches in panel data
- Other directions previously studied
  - Stockpiling, learning/experimentation, habit formation, effects of advertising/coupons/promotions

Existing Approaches: CS

Typical approach from the computer science literature:

- Canonical example: Netflix movie recommendations
- Estimating correlations in preferences between customers, ignoring substitutes/complements



"What types of things do customers like?"

Asking and Answering Questions Using Panel Data with Consumer Choices Over Many Products

- Our Approach
  - Build on matrix factorization and embedding methods from CS
  - Use Bayesian approach for flexibility in incorporating structure
  - Estimate a structural model
  - Pay attention to identification and supplementary analyses in environment with many small experiments
- Two Ways to Use the Results
  - Traditional structural model: counterfactuals within the model
  - As a pre-processing step–to reduce the dimensionality and gain efficiency
    - Infer user preferences for rarely purchased products, rather than fixed effects
    - Motivate functional form assumptions–which products are potentially substitutes, complements, or independent
    - Reduce dimensionality of outcome space–e.g. Rover.com customers consumer more in latent categories related to travel

Towards A Large Scale Model of Consumer Choice

Goals of this agenda:

- Apply large-scale latent variable approaches with multiple unobserved product characteristics and latent user preferences over items (Poisson factorization: Gopalan, Hofman, Blei (2013); Gopalan, Charlan, Blei (2014))
- Consider many products at once
- Distinguish correlation from complementarity (price/availability change over time)
- Test assumptions and validate causal models
- Counterfactuals: consumer welfare for different pricing policies
- Evaluate the effects of policy changes, product introductions, or shocks to consumers

Steps

1. Validate price identification strategy systematically
2. Item by item choice: unified model v. separate by category
3. Complements

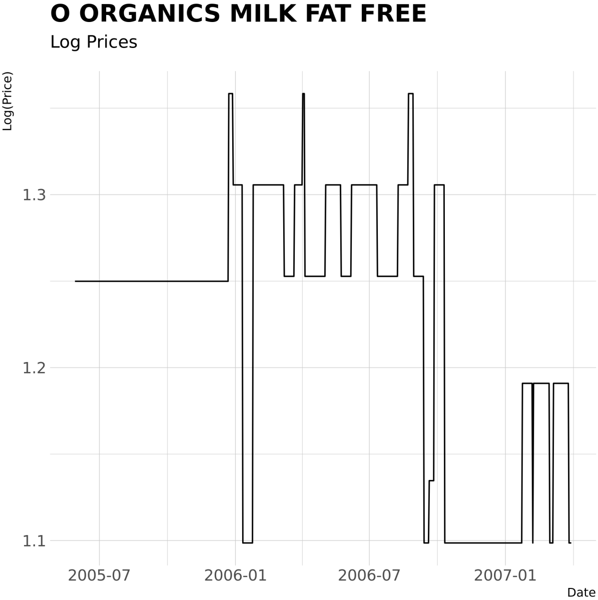
The Data

- Dataset constructed by Che, Chen and Chen (2012)
- All loyalty-card shoppers at a single, isolated store over 18 month period
- Out of stock data at approximately hourly level (daily is good enough)
- Almost all prices change on Tuesday night; focus on Tuesday and Wednesday data
- Product hierarchy (UPC, subclass, class, category, group, department section, department)
- User demographics: we include 28 variables derived from a variety of sources, may have measurement error

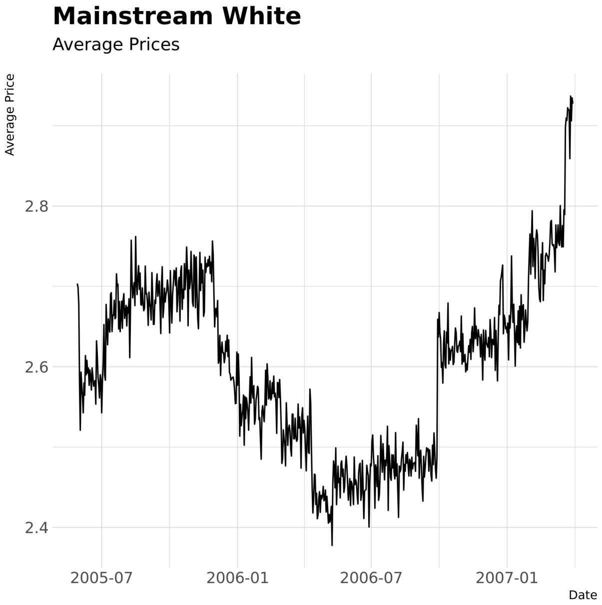
The Data: Sample Selection

- Users who had at least 20 shopping trips on Tuesday or Wednesday with more than 10 items per trip
- Top 235 categories
- Further restrictions
  - More than one UPC in category
  - More than one top ten product w/ price variation
  - Purchases not too concentrated
  - Prices not too highly correlated
  - Less than 10% buy multiple items per purchase in category
  - More than one top 10 UPC with price changes greater than $.10 in at least 10% of weeks
- Dataset: 2068 consumers, 123 categories, 1263 items, 333,585 trips
- Training/Validation/Test: 65%, 5%, 30%

A Sample UPC-Level Price Series



A Sample Category-Level Price Series



Assumptions: Discussion

Our "identifying assumption" is:

- Counterfactual distributions of purchases between Tues & Wed in weeks with price changes can be constructed from behavior on weeks without price changes
- Weeks with price changes, and magnitudes of price changes, have the same day of week/time of day patterns as other weeks

This assumption would be violated if, for example:

- Store chooses to lower prices in a week where demand would have been growing through the week (e.g. due to minor holiday at end of week)
- Prices generally trending up or down in the sample, corresponding to decline or increase in a product's popularity, marketing, fruit coming into season, etc.

Testing Assumptions: Pseudo-Treatments

If prices and quantities both have systematic time trends, fixed effect models will not fully control for them

- Include global time trends as controls (no effect in most products)
- Placebo test
  - Shift price series forward (or backward) in time
  - Skip over weeks that also have a price change
  - Re-estimate model, calculate p-values

Supplementary Analysis

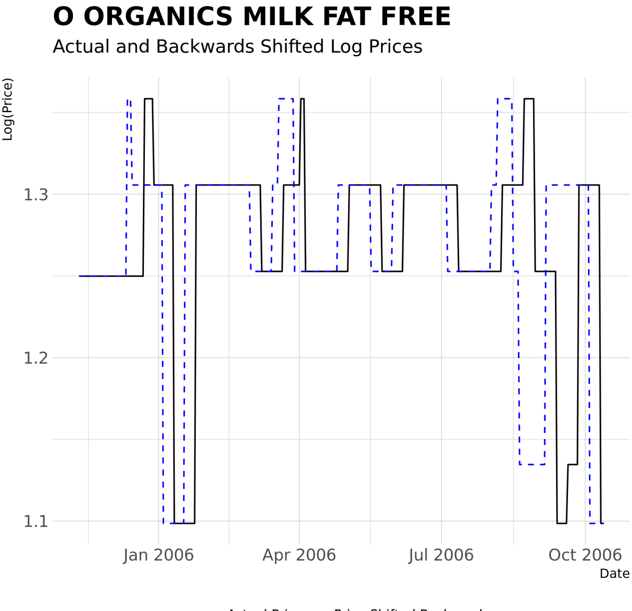
In simple model, do variants of "Placebo tests" to see if our identification strategy is valid.

- Use only shopping trips t on Tuesday and Wednesday
- Estimate multinomial logit model with

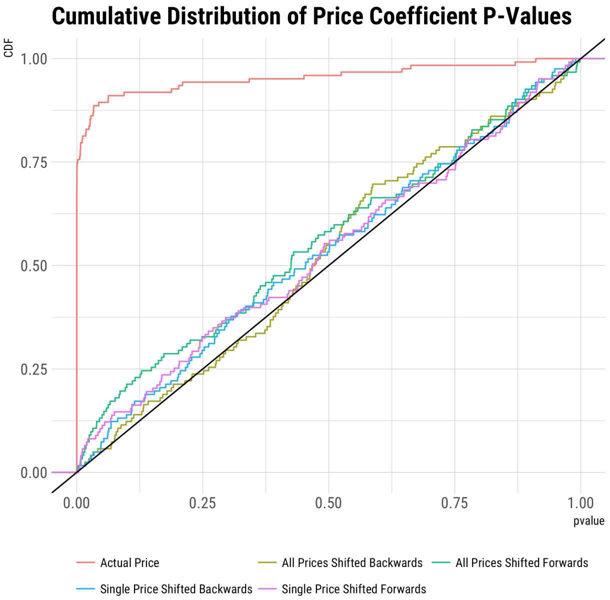
$$U_{uit} = \alpha \log(P_{it}) + X_{it}\beta' + \epsilon_{uit} \quad (1)$$

- $\log(P_{it})$ log price of item during trip
- X_{it} contains
  - Tuesday and Wednesday dummies
  - Week pseudo-fixed effects (week average category purchases)

A Sample UPC-Level Price Series



Placebo Test Results



The Baseline Poisson Factorization (HPF) Model

- User u has K-vector of non-negative preferences θ_u
- Item i has a K-vector of non-negative attributes β_i.
- Utility for item

$$U_{uit} = \log(\theta_u^\top \beta_i) + \epsilon_{uit} \tag{2}$$

  where ϵ_{uit} is drawn from extreme value distribution.
- She chooses to buy if utility is positive.
- Choices about products are independent
- Parameters are all positive

The Nested Logit Factorization Model

- User u has K-vector of preferences θ_u and a vector of price sensitivity parameters γ_u
- Item i has two K-vector of attributes α_i and β_i.
- Utility for item

$$U_{uit} = \theta_u^\top \beta_i - \gamma_u^\top \alpha_i ln(p_{uit}) + \epsilon_{uit} \tag{3}$$

  where ϵ_{uit} is drawn from extreme value distribution and are independent conditional on purchasing an item within a category.
- She chooses the highest utility item in each category or the outside option; the outside option is in its own nest
- Choices about categories are independent

The Nested Logit Factorization Model Cont'd

- Users u independently chooses whether or not to make a purchase from each product category c.
- Utility for not choosing category

$$U_{uc_0t} = \theta_{c,u}^{\top} \beta_{c_0} + \epsilon_{uc_0t} \tag{4}$$

- Utility for choosing category

$$U_{uc_1t} = \theta_{c,u}^{\top} \beta_{c_1} - \mu_u \delta_{c_1} IV_c + \epsilon_{uc_1t} \tag{5}$$

- Where IV_c is the inclusive value of the items in the category is given by $IV_c = \log \sum_{i \in J_c} \exp U_{uit}$, which is the expectation of the max of the U_{uit} prior to learning the ϵ_{uit} for each item.

Estimation of Model

- MCMC-based Bayesian methods: Common in marketing for estimating models with heterogeneity, but computationally infeasible as data size and number of parameters grows
- This choice of functional form allows for fast and efficient estimation using variational Bayesian inference
- Variational Bayes:
  - Choose parameterized family of distributions $q(\cdot|\eta)$ to approximate the posterior
  - Find η that minimizes KL-divergence to the true posterior
  - With appropriate choice of priors and q, this optimization can be done using simple coordinate ascent
  - Accuracy similar to MCMC, but 1000s of times more quickly
- Introducing price effects and time-varying price slows things down substantially (hours rather than minutes; but still feasible unlike MCMC)
- Introducing substitutability within categories requires additional computational tricks

Model Details

- Item characteristics: Department section indicator variables, Price
- Customer characteristics: Gender, Age Bracket, Marital Status, Children, Income Bracket
- Control for weekly average purchases at the category level
- Number of latent characteristics chosen through validation

Model Comparisons

- Main benefits of our model: personalization and efficiency from pooling categories
- Can compare to standard models (MNL, nested logit, mixed logit)
- Can also look at benefit from taking HPF estimates of personalized mean utility for each item, and including these in standard models
- Note bias-variance tradeoff issues:
  - If you evaluate at individual level, better to have model with high personalization but some bias; at aggregate level, need to eliminate bias with simpler model that gets average right

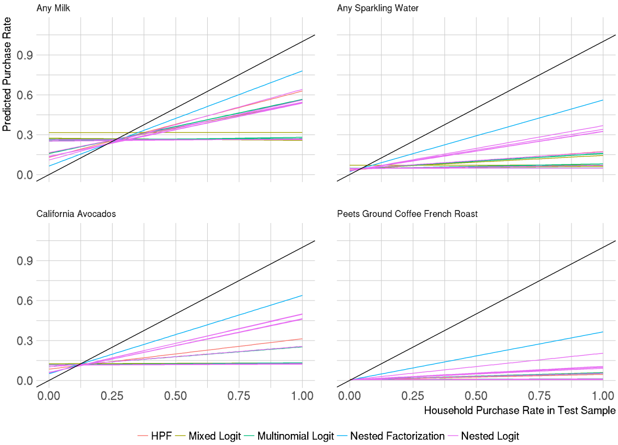
Table: Predictive Performance (Test Set Mean Log Likelihood)

	Mean Log Likelihood		Mean Squared Error	
Model	Train	Test	Train	Test
Nested Factorization	-4.1776	-4.9651	0.8911	0.9313
Multinomial Logit with Item-Specific HPF	-5.2264	-5.4108	0.9554	0.9619
Multinomial Logit with Frequency x HPF Interaction	-5.2268	-5.4124	0.9556	0.9621
Multinomial Logit with HPF Controls	-5.2285	-5.4146	0.9557	0.9622
Mixed Logit with Random Price Effects and HPF Controls	-5.2403	-5.4221	0.9561	0.9624
Nested Logit with HPF Controls	-5.2480	-5.4231	0.9560	0.9623
Hierarchical Poisson Factorization (HPF)	-5.2698	-5.4643	0.9516	0.9582
Multinomial Logit with Demographic and Frequency Controls	-5.6213	-5.6781	0.9784	0.9792
Nested Logit with Demographic Controls	-5.6192	-5.6798	0.9782	0.9792
Multinomial Logit with Behavioral and Frequency Controls	-5.6333	-5.6815	0.9786	0.9793
Multinomial Logit with Demographic Controls	-5.6249	-5.6827	0.9785	0.9794
Nested Logit with Behavioral Controls	-5.6400	-5.6922	0.9787	0.9795
Multinomial Logit with Behavioral Controls	-5.6456	-5.6952	0.9789	0.9797
Multinomial Logit with Frequency Dummy	-5.6808	-5.7127	0.9795	0.9800
Nested Logit	-5.6787	-5.7141	0.9794	0.9800
Multinomial Logit	-5.6842	-5.7170	0.9796	0.9801
Mixed Logit with Random Price Effects and Demographics	-5.6924	-5.7272	0.9798	0.9804
Mixed Logit with Random Price Effects and Behavioral Controls	-5.6843	-5.7320	0.9798	0.9804
Mixed Logit with Random Price Effects	-5.7055	-5.7361	0.9800	0.9805
Mixed Logit with Random Price and Random Intercepts	-5.8043	-5.8330	0.9826	0.9830
Nested Logit with Category Spending and HPF Controls	-4.6383	-6.5533	0.9357	0.9520
Nested Logit with Demographic and Category Spending Controls	-4.9004	-6.7376	0.9602	0.9709
Nested Logit with Behavioral and Category Spending Controls	-4.9091	-6.7481	0.9607	0.9715
Nested Logit with Category Spending Controls	-4.9451	-6.7610	0.9626	0.9723

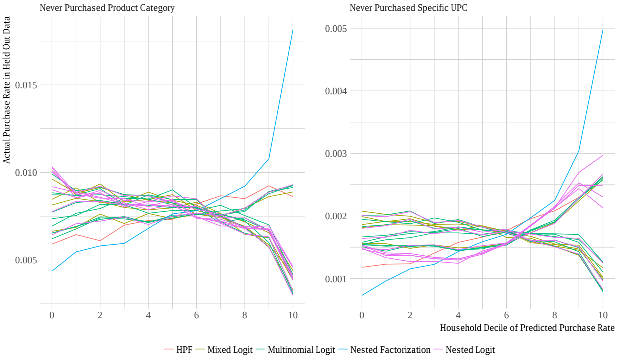
Table: Counterfactual Performance (Test Set Mean Log Likelihood)

Model	All Weeks	Out of Stock	Own Price Change	Cross Price Change
Nested Factorization	-3.884 (0.021)	-2.925 (0.041)	-5.550 (0.058)	-3.315 (0.034)
Multinomial Logit with Item-Specific HPF	-4.233 (0.022)	-3.200 (0.044)	-5.968 (0.060)	-3.627 (0.037)
Multinomial Logit with Frequency x HPF Interaction	-4.234 (0.022)	-3.200 (0.044)	-5.968 (0.060)	-3.629 (0.037)
Multinomial Logit with HPF Controls	-4.236 (0.022)	-3.202 (0.044)	-5.970 (0.060)	-3.630 (0.037)
Mixed Logit with Random Price Effects and HPF Controls	-4.242 (0.022)	-3.209 (0.044)	-5.978 (0.060)	-3.636 (0.037)
Nested Logit with HPF Controls	-4.243 (0.022)	-3.210 (0.044)	-5.990 (0.060)	-3.640 (0.037)
Hierarchical Poisson Factorization (HPF)	-4.275 (0.022)	-3.270 (0.046)	-6.072 (0.062)	-3.642 (0.037)
Multinomial Logit with Demographic and Frequency Controls	-4.442 (0.023)	-3.326 (0.045)	-6.166 (0.061)	-3.839 (0.039)
Nested Logit with Demographic Controls	-4.443 (0.023)	-3.329 (0.045)	-6.167 (0.061)	-3.841 (0.039)
Multinomial Logit with Behavioral and Frequency Controls	-4.445 (0.023)	-3.328 (0.045)	-6.168 (0.062)	-3.840 (0.039)
Multinomial Logit with Demographic Controls	-4.446 (0.023)	-3.330 (0.045)	-6.172 (0.062)	-3.841 (0.039)
Nested Logit with Behavioral Controls	-4.453 (0.023)	-3.337 (0.045)	-6.178 (0.062)	-3.848 (0.040)
Multinomial Logit with Behavioral Controls	-4.455 (0.023)	-3.339 (0.045)	-6.183 (0.062)	-3.849 (0.039)
Multinomial Logit with Frequency Dummy	-4.469 (0.023)	-3.351 (0.045)	-6.194 (0.062)	-3.863 (0.040)
Nested Logit	-4.470 (0.023)	-3.352 (0.045)	-6.195 (0.062)	-3.865 (0.040)
Multinomial Logit	-4.472 (0.023)	-3.355 (0.046)	-6.200 (0.062)	-3.866 (0.040)
Mixed Logit with Random Price Effects and Demographics	-4.480 (0.023)	-3.357 (0.045)	-6.200 (0.062)	-3.881 (0.039)
Mixed Logit with Random Price Effects and Behavioral Controls	-4.484 (0.023)	-3.360 (0.045)	-6.209 (0.062)	-3.886 (0.039)
Mixed Logit with Random Price Effects	-4.487 (0.023)	-3.366 (0.045)	-6.211 (0.062)	-3.885 (0.039)
Mixed Logit with Random Price and Random Intercepts	-4.563 (0.023)	-3.444 (0.044)	-6.270 (0.061)	-3.965 (0.039)
Nested Logit with Category Spending and HPF Controls	-5.127 (0.029)	-3.948 (0.065)	-7.150 (0.078)	-4.277 (0.049)
Nested Logit with Demographic and Category Spending Controls	-5.271 (0.030)	-4.011 (0.065)	-7.268 (0.079)	-4.438 (0.050)
Nested Logit with Behavioral and Category Spending Controls	-5.279 (0.030)	-4.020 (0.065)	-7.269 (0.079)	-4.448 (0.051)
Nested Logit with Category Spending Controls	-5.289 (0.030)	-4.023 (0.066)	-7.285 (0.079)	-4.457 (0.051)

Actual v. Predicted Across Models



Never Purchased Categories and Items



Elasticities: Within Class v. Across Classes

- Compare within-class v. across-class elasticities; and within subclass and across-subclass elasticities.
- Nested Logits give 10% higher within-class than across-class
- MNL approximately the same within and across.
- Nested factorization model gives 6=70% higher within-class. (preliminary magnitudes)

Using Estimates

- Elasticities used for pricing and personalized coupon strategy
- Put users into clusters for tracking, policy assignment
- Use estimated parameters as observables in structural model
  - Pooling information across users and products can increase efficiency in environment with many products and users
  - Use predicted user-product mean utilities (including user and item observables) in place of user-product fixed effects in MNL model

Mean Utility v. Elasticity

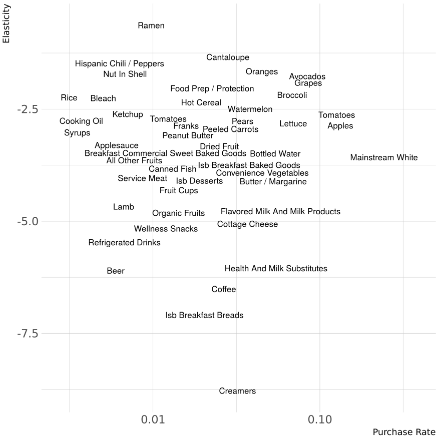


Figure: Comparison of Predicted Purchase Probabilties and Price Elasticities Across Product Categories (median values across all household x item pairs)

Substitutes and Complements

$$U_{ubt} = \mu_{ub} + \epsilon_{ubt}$$

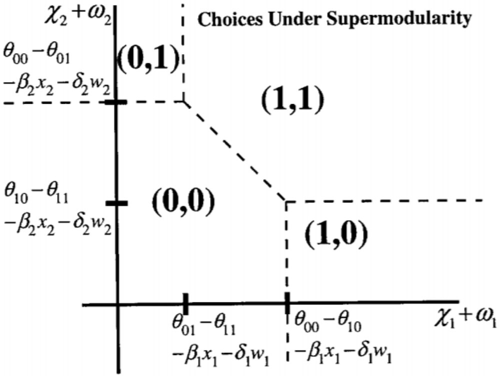
OR

$$U_{ubt} = \mu_{ub} + \sum_{i \in B} \epsilon_{uit}$$

- Can consumer really maximize utility over $B = 2^I$ bundles when $I = 5000$ or more?
- Can we as econometricians estimate such a model?
  - Athey and Stern (1998): This is hard problem computationally with $I = 3$ or 4, even with assns and factorization (item-specific independent errors)

Substitutes and Complements

FIGURE 1



Athey-Stern discuss identification through variation in X, W (like prices), and estimation through numerical integration of the unobservables

Substitutes and Complements

- Surprisingly little literature on empirical estimation since Athey and Stern (1998), and almost all have 2-3 choices
- Gentzkow (2007) implements similar approach
- Berry et al (2014), Chintagunta and Nair (2011) survey
- Train, McFadden and Ben-Akiva (1987) treat each bundle as a discrete alternative, but use nested logit to account for correlation among related bundles
- Song and Chintagunta (2007) build a utility-maximization framework where consumers select not just whether to purchase, but how much, and apply it to supermarket purchase data for two products, laundry detergent and fabric softener.
- See also Seiler, Thomassen, Smith and Schiraldi (2017); Smith, Rossi, and Allenby (2017); Wan, Wang, Goldman, Taddy, Rao (2017)

Consumer Shopping Heuristic

Behavioral assumption

- Assume myopic consumer goes into the store and considers sequentially what to buy
- Selects the item that maximizes utility myopically (for now: over whole store)–assuming will not buy anything else–but accounts for complementarity with what is already in basket
- "Look-ahead model": look ahead to the next item (for now: over whole store) when making choice, accounting for potential complementarity

Estimation: latent item orderings

- For each order, assume analyst observed that order
- Model implies a likelihood over each ordering (higher value items purchased earlier)

Simulation
Comparing Myopic to Look-Ahead Models

		stage 1: Diapers	stage 2: Hotdogs	stage 3: Buns	stage 4: *checkout*
non think-ahead	diapers	**0.39**	0.00	0.00	0.00
	coffee (↑)	0.01	0.02	0.03	0.12
	ramen	0.00	0.00	0.00	0.00
	candy	0.00	0.00	0.00	0.00
	hot dogs	0.18	**0.20**	0.00	0.00
	hot dog buns	0.18	0.24	**0.79**	0.00
	taco shells (↑)	0.03	0.07	0.00	0.00
	taco seasoning	0.21	0.41	0.00	0.00
	checkout	0.00	0.06	0.18	**0.88**
think-ahead	diapers	**0.38**	0.00	0.00	0.00
	coffee (↑)	0.02	0.01	0.06	0.05
	ramen	0.00	0.00	0.00	0.00
	candy	0.00	0.00	0.00	0.00
	hot dogs	0.23	**0.40**	0.00	0.00
	hot dog buns	0.32	0.53	**0.87**	0.00
	taco shells (↑)	0.02	0.02	0.00	0.00
	taco seasoning	0.02	0.03	0.00	0.00
	checkout	0.00	0.01	0.06	**0.95**

Goodness of Fit

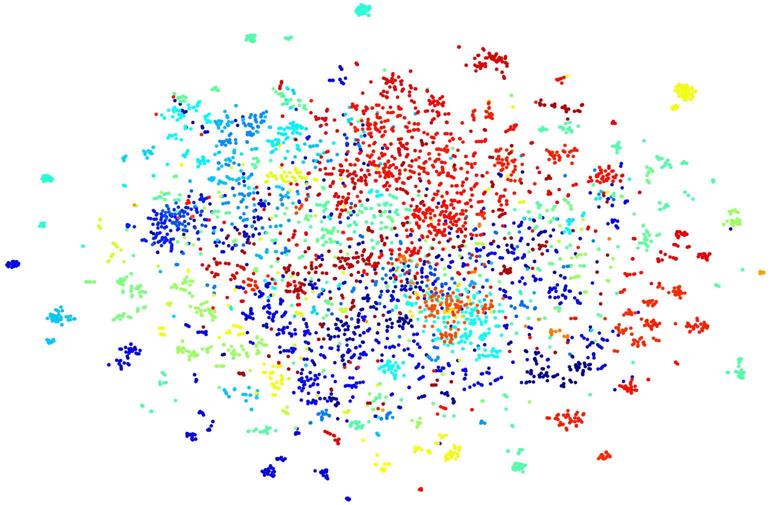
Model	Log-likelihood All (320K)	Price±10% (66K)	Price±20% (20K)	Price±30% (1.5K)
B-Emb	−5.13	−5.30	−5.33	−5.35
P-Emb	−5.13	−5.34	−5.42	−5.48
HPF	−4.97	−5.24	−5.35	−5.45
This paper (I+U)	−4.94	−5.21	−5.27	−5.33
This paper (I+U+P)	−4.93	−5.13	−5.09	−5.01
This paper (I+U+P+S)	**−4.92**	**−5.12**	**−5.08**	**−5.00**

Table: Average predictive log-likelihood on the test set, conditioning on the remaining items of each basket. SHOPPER with user preferences improves over the existing models. The improvement grows when adjusting for price and seasonal effects, and especially so when using skewed test sets that emulate price intervention.

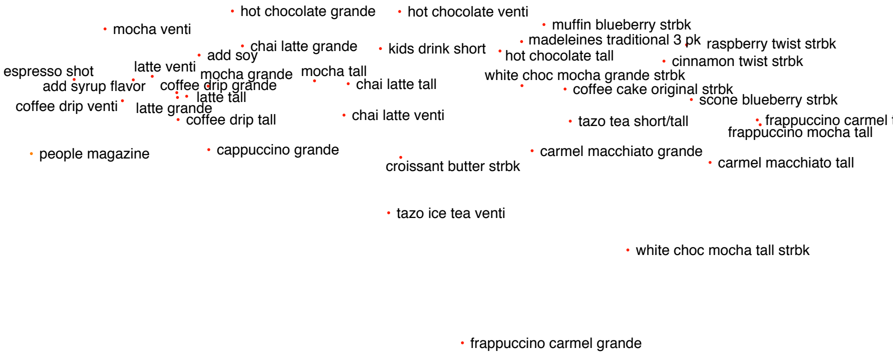
Role of Lookahead Model

	Three items	Entire baskets
Non think-ahead	-4.93	-5.14
Think-ahead	$-\mathbf{4.82}$	$-\mathbf{5.05}$

Representation of Latent Item Characteristics

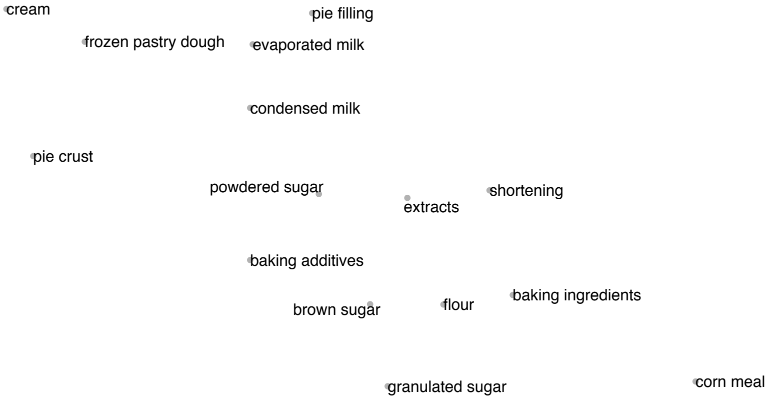


UPCs that are Close in α space

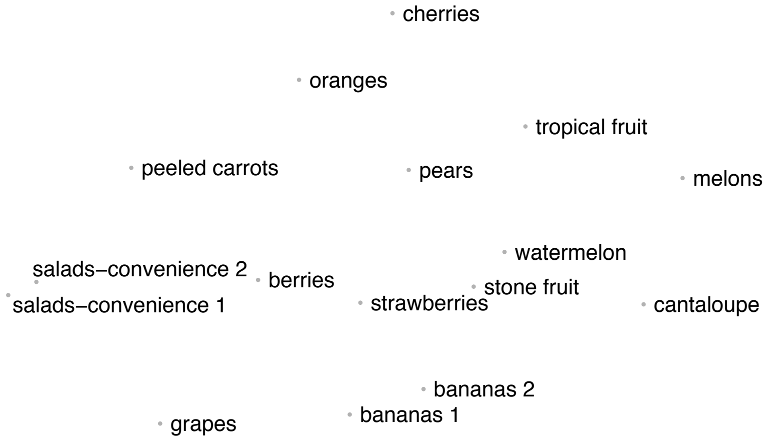


UPCs that are Close in α space

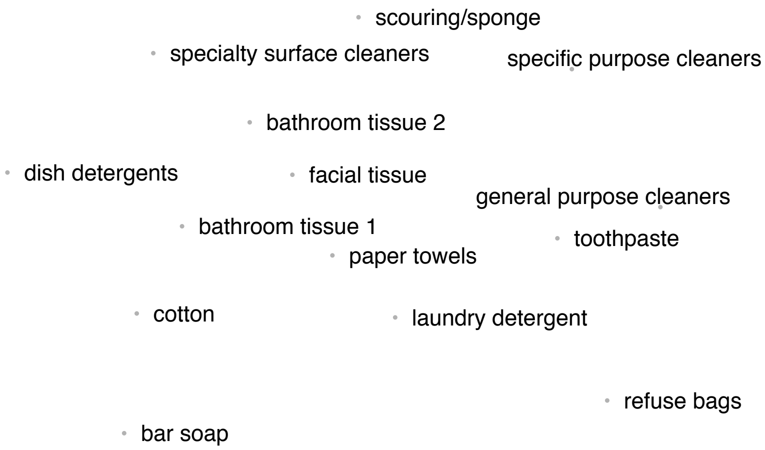
Categories that are Close in α space



Categories that are Close in α space



Categories that are Close in α space



Similarity/Exchangeability v. Complementarity

query items	complementarity score		exchangeability score	
mission tortilla soft taco	2.51	ortega taco shells white corn	0.05	mission fajita size
	2.40	mcrmck seasoning mix taco	0.10	mission tortilla fluffy gordita
	2.26	lawrys taco seasoning mix	0.11	mission tortilla soft taco
private brand hot dog buns	3.02	bp franks bun size	0.10	private brand hamburger buns
	2.94	bp franks beef bun length	0.12	ball park buns hot dog
	2.86	private brand hamburger buns	0.14	private brand hot dog buns ssme 8ct
private brand mustard squeeze bottle	0.53	private brand hamburger buns	0.14	frenchs mustard classic yellow squeeze
	0.44	private brand cutlery full size asst	0.16	frenchs mustard classic yellow squeezed
	0.29	private brand hot dog buns	0.17	heinz ketchup squeeze bottle
private brand napkins all occasion	1.01	private brand cutlery full size forks	0.08	vnty fair napkins all occasion
	0.62	dixie heavy duty plates dspbl 10 1/4 in	0.10	vnty fair napkins all occasion
	0.39	private brand plate dsgnr 6 7/8 in	0.13	glad cling wrap plastic wrap

Table: Items with the highest complementarity and lowest exchangeability metrics for some query items.

Conclusions

- Factorization is effective technique that enables personalization with dimension reduction
- Bayesian/structural models enable incorporation of functional forms motivated by theory and practice in economics
- Important to tune models for desired goals (counterfactuals)
- Models can discover product interaction and price sensitivities with minimal external information about products
- These models are a baseline against which we can proceed to estimate impact of other types of interventions
- In progress: applications to movement data