Looking into the Consumption Black Box:

Evidence from Scanner Data

Krzysztof Pytka (Mannheim) ⓡ Daniel Runge (ECB)

University of Edinburgh, 9/25/25

Generative prompt:

```
Create a painting entitled
"Looking into the Consumption Black Box"
painted by Sandro Botticelli
```

Disclaimers

Data Disclaimer

The researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Affiliation Disclaimer

This paper should not be reported as representing the views of the European Central Bank.
The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project

"There is no accounting for taste"

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project

"There is ~~no~~ accounting for taste"

◇ **Neoclassical model with representative agent**

(Ramsey, EJ 1928; Cass, REStud 1965; Koopmans, 1965; Brock&Mirman, JET 1972; Kydland&Prescott, Ecta 1982)

- Early macroeconomic models considered a single representative agent.
- Consumption was aggregated into a single product for tractability.
- This approach ignored individual differences in consumption patterns.

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project

"There is ~~no~~ accounting for taste"

- **Neoclassical model with representative agent**

  (Ramsey, EJ 1928; Cass, REStud 1965; Koopmans, 1965; Brock&Mirman, JET 1972; Kydland&Prescott, Ecta 1982)

  - Early macroeconomic models considered a single representative agent.
  - Consumption was aggregated into a single product for tractability.
  - This approach ignored individual differences in consumption patterns.

- **Models of Inequalities w/Heterogeneous Households**

  (Aiyagari, QJE 1994, JPE 1995; Huggett, JEDC 1993)

  - Aimed to capture state-based consumption disparities.
  - Generate differences in quantities consumed across households.
  - Still focused on a single type of product.

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project

"There is ~~no~~ accounting for taste"

- **Neoclassical model with representative agent**

  (Ramsey, EJ 1928; Cass, REStud 1965; Koopmans, 1965; Brock&Mirman, JET 1972; Kydland&Prescott, Ecta 1982)

  - Early macroeconomic models considered a single representative agent.
  - Consumption was aggregated into a single product for tractability.
  - This approach ignored individual differences in consumption patterns.

- **Models of Inequalities w/Heterogeneous Households**

  (Aiyagari, QJE 1994, JPE 1995; Huggett, JEDC 1993)

  - Aimed to capture state-based consumption disparities.
  - Generate differences in quantities consumed across households.
  - Still focused on a single type of product.

- **Models of Heterogeneous Goods (+Heterogeneity in Choices)**

  (Handbury, Ecta 2021; Michelacci, Paciello, and Pozzi, REStud 2022; Neiman&Vavra, AEJ:Macro 2023; Mongey&Waugh, JPE R&R 2025; Bornstein&Peter, 2025)

  - Recognize that households may consume different varieties of products due to: either (i) (intratemporal) non-homothetic preferences or (ii) latent heterogeneity.
  - Differences in preferences lead to diverse consumption baskets.
  - Aim to provide more realistic representation of actual consumption behavior.

Testable Implications of Models w/Heterog. Goods

Non-Homothetic Preferences

- Richer households shift toward different products
- Expenditure shares change with income

Latent Heterogeneity

- Similar households have different baskets
- Choices made by individual buyers should be rather stable over time

Testable Implications of Models w/Heterog. Goods

Non-Homothetic Preferences

- Richer households shift toward different products
- Expenditure shares change with income

Latent Heterogeneity

- Similar households have different baskets
- Choices made by individual buyers should be rather stable over time

Empirical Investigation

These models have testable implications that can be empirically studied.

Research Questions and Why They Matter

Research Questions ?

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?

Research Questions and Why They Matter

Research Questions ?

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?
3. What are implications for structural models of the answers to Questions 1 and 2?

Research Questions and Why They Matter (*cont'd*)

- While there is a rich empirical literature documenting consumption inequality, much less is known about what different groups actually buy

Research Questions and Why They Matter (*cont'd*)

- While there is a rich empirical literature documenting consumption inequality, much less is known about what different groups actually buy
- This research gap has **direct policy implications** for areas such as:
  - **Price discrimination strategies by retailers**
    (Mussa & Rosen, JET 1978; Maskin & Riley, RAND 1984)
  - **Redistribution through commodity taxation**
    (Saez, JPubE 2003; Atkinson & Stiglitz, JPubE 1976)
  - **Magnitude of consumption-search externalities**
    (Pytka, REStud R&R 2024)

Research Questions and Why They Matter (*cont'd*)

- While there is a rich empirical literature documenting consumption inequality, much less is known about what different groups actually buy

- This research gap has **direct policy implications** for areas such as:
  - **Price discrimination strategies by retailers**
    (Mussa & Rosen, JET 1978; Maskin & Riley, RAND 1984)
  - **Redistribution through commodity taxation**
    (Saez, JPubE 2003; Atkinson & Stiglitz, JPubE 1976)
  - **Magnitude of consumption-search externalities**
    (Pytka, REStud R&R 2024)

Key Research Challenges

1. **Conceptual/computational:**
   How to measure group differences in high-dimensional consumption choices?
2. **Statistical:**
   Addressing potentially severe small-sample bias in finite samples (yet to be discussed)

Preview of the Results

Limited Consumption Polarization

- Rich and poor households show surprisingly similar spending patterns
- On average, observing how a consumer spends $1 provides very modest information about their income level
- We document data sparsity that leads to overestimated polarization if unaddressed

Highly Unstable Individual Choices

- Only 40% of products repurchased year-to-year

Parsimonious Model (*Shopping Spree*)

- Perfect substitutes + random sampling fits data well
- But yields fundamentally different policy implications than heterogeneous preference models
- **Identification problem:** if two very different models fit equally well, how do we know which mechanisms drive consumption patterns?

Literature Review

- **Heterogeneous Goods**
  Handbury (Ecta, 2021), Neiman & Vavra (AEJ:Macro, 2023),
  Michelacci, Paciello, & Pozzi (REStud, 2022), Faber & Fally (REStud, 2022),
  Becker (JMP, 2024), Mongey & Waugh (JPE R&R, 2025), Mangin (Ecta, *Forth.*),
  Bornstein & Peter (2025)

- **Polarization in Economics**
  Bertrand & Kamenica (AEJ: Applied, 2023), Alesina, Tabellini, & Trebbi (BP:EA, 2017),
  Desmet & Wacziarg (EJ, 2021), Boar & Giannone (WP, 2023)

- **Big Data with Small Sample Bias**
  Gentzkow, Shapiro, & Taddy (Ecta, 2019), Armenter & Koren (AER, 2014)

- **Methodological Debates** – *"Stirring up a hornet's nest"*
  - *Confronting Mainstream:* Armenter & Koren (AER, 2014), Menzio (JPE R&R, 2025), Borovičková & Shimer (QJE R&R, 2024)
  - *Defending Mainstream:* Becker (JPE, 1962)

Data: NielsenIQ Consumer Panel

- **Source:** Kilts-NielsenIQ Consumer Panel
- **Sample Coverage:** 40,000–60,000 households per year (2004–2016) with detailed socio-economic characteristics
- **Product Universe:** 1.5 million unique products ▪ groceries ▪ drug products ▪ small appliances ▪ electronics
- **Data Quality:** Scanner-based collection ensures low measurement error
  - Data collected via handheld scanner or mobile app by scanning barcodes
- **Representativeness:** Projection weights ensure US economy representativeness

Main Questions

Research Questions ?

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?
3. What are implications for structural models of the answers to Questions 1 and 2?

Polarization in Consumption

Challenge 1

How to measure group differences in **high-dimensional** choices?

- Measuring inequality is dramatically simpler: Gini coefficient, Lorenz curve, etc.
- Measuring polarization is challenging due to high-dimensional data and diverse product categories. ▸ Some products are favored by specific income groups ▸ Other products have universal appeal

Polarization in Consumption

Challenge 1

How to measure group differences in **high-dimensional** choices?

- Measuring inequality is dramatically simpler: Gini coefficient, Lorenz curve, etc.
- Measuring polarization is challenging due to high-dimensional data and diverse product categories. ▸ Some products are favored by specific income groups ▸ Other products have universal appeal

Our Approach: Gentzkow, Shapiro, & Taddy (Ecta, 2019) ▸ Methodological Context

How much can I learn about your income from what you buy?

Testable Implication

If preferences are non-homothetic, purchases reveal income information

How much can I learn about your income from what you buy?

Testable Implication

If preferences are non-homothetic, purchases reveal income information

$q_t^G(j) :=$ probability that \$1 spent by group G goes to product j

Groups:

- **High-expenditure (H):** top quintile
- **Low-expenditure (L):** bottom quintile

How much can I learn about your income from what you buy?

Testable Implication

If preferences are non-homothetic, purchases reveal income information

$q_t^G(j) :=$ probability that \$1 spent by group G goes to product j

Groups:

- **High-expenditure (H):** top quintile
- **Low-expenditure (L):** bottom quintile

Non-homotheticity predicts: $q_t^H(j) \neq q_t^L(j)$

How much can I learn about your income from what you buy?

Testable Implication

If preferences are non-homothetic, purchases reveal income information

$q_t^G(j) :=$ probability that \$1 spent by group G goes to product j

Groups:

- **High-expenditure (H):** top quintile
- **Low-expenditure (L):** bottom quintile

Non-homotheticity predicts: $q_t^H(j) \neq q_t^L(j)$

Question

How much income information do spending patterns contain?

Bayesian Learning from Purchases

After observing someone purchase product j, we update our beliefs:

Posterior probability of high income:

$$\rho_t(j) = P(H \mid j) = \frac{q_t^H(j)}{q_t^H(j) + q_t^L(j)}$$

Posterior probability of low income:

$$P(L \mid j) = 1 - \rho_t(j)$$

Key insight: The informativeness of observing purchase j depends on the ratio of conditional probabilities:

- If $q_t^H(j) \gg q_t^L(j)$: observing purchase j strongly suggests high income
- If $q_t^H(j) \ll q_t^L(j)$: observing purchase j strongly suggests low income
- If $q_t^H(j) \approx q_t^L(j)$: the purchase is uninformative about income

Examples from the NielsenIQ universe: Computing $\rho_t(j)$

Soft'n Gentle Tissue

Purchased **1.98× more** by low-income consumers

Tropicana Pure Premium

Purchased **7.56× more** by high-income consumers

Kellogg's All-Bran

Purchased at **similar rates** by both groups

$\rho(\text{Soft'n Gentle}) = 0.34$

$\rho(\text{Tropicana}) = 0.89$

$\rho(\text{Kellogg}) = 0.50$

Strong signal of **low income**

Strong signal of **high income**

Uninformative about income

Pattern: Products with extreme purchase ratios between income groups yield ρ_j values close to 0 or 1, making them highly informative for income prediction.

Measuring Prediction Accuracy

Consumption-based polarization (*aka* Bayesian accuracy of a signal)

$$\bar{\pi}_t = \frac{1}{2}\sum_j q_t^H(j)\cdot\rho_t(j) + \frac{1}{2}\sum_j q_t^L(j)\cdot[1-\rho_t(j)]$$

Methodology Summary

1. Randomly select $1 spent in Nielsen universe
2. Observe product allocation of this $1
3. Predict household group membership from products
4. **Polarization** = expected predictive accuracy

Interpretation

No polarization $\xrightarrow{\text{Increasing predictive power}}$ Perfect polarization

0.5 ———— **1.0**

Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. **50,000** households $\Rightarrow$ **Sparse Data Problem**

Big Data with Small-Sample Bias

Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. **50,000** households ⇒ **Sparse Data Problem**

View Table

- **Data Sparsity:** ≈ **30%** of consumption bought only once
- **Extreme Case:** Single purchases ⇒ 100% predictive power
- **Key Question:** Are rare purchases systematic or random?

Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. **50,000** households $\Rightarrow$ **Sparse Data Problem**

- **Data Sparsity:** $\approx$ **30%** of consumption bought only once

  View Table
- **Extreme Case:** Single purchases $\Rightarrow$ 100% predictive power
- **Key Question:** Are rare purchases systematic or random?

Solution Strategy

- **Cannot use raw data** $\Rightarrow$ infrequent transactions cause spurious polarization
- Need to **filter systematic differences** and **reduce noise**
- **Solution:** Estimate theoretical distributions, then study polarization

Measuring Consumption Polarization

Challenge: Small sample bias when comparing raw consumption patterns

Measuring Consumption Polarization

Challenge: Small sample bias when comparing raw consumption patterns

Solution: Following Gentzkow, Shapiro, & Taddy (Ecta, 2019)

- Estimate consumption choice model
- Compare predicted (not raw) choice distributions
- Use regularization to extract systematic differences

In a nutshell: Study estimated consumption generating processes, not raw data

Model Specification

Household consumption:

$$\mathbf{c}_{i,t} \sim \text{Multinomial}(m_{i,t}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{i,t}))$$

where:

- $m_{i,t}$ = total expenditure
- $P(i) \in \{H, L\}$ = income group
- $\mathbf{x}_{i,t}$ = household characteristics

Choice probabilities: Multinomial logit

$$q_{jt}^{P(i)}(\mathbf{x}_{i,t}) = \frac{e^{u_{i,j,t}}}{\sum_{k \in J} e^{u_{i,k,t}}}$$

Identifying Income Group Differences

Choice probabilities:

$$q_{jt}^{P(i)}(\mathbf{x}_{i,t}) = \frac{e^{u_{i,j,t}}}{\sum_{k \in J} e^{u_{i,k,t}}}$$

$$u_{i,j,t} = \alpha_{j,t} + \mathbf{x}_{i,t}'\gamma_{j,t} + \varphi_{j,t}\mathbb{1}_{i \in H_t}$$

- $\alpha_{j,t}$ = product-year fixed effects
- $\gamma_{j,t}$ = demographic controls
- $\varphi_{j,t}$ = income group difference

Key parameter: $\varphi_{j,t}$ captures how high vs. low income groups differ

LASSO Regularization

Problem: 800,000+ product categories $\rightarrow$ overfitting risk

LASSO Regularization

Problem: 800,000+ product categories $\rightarrow$ overfitting risk

Solution: Penalize income group differences

$$\sum_{j,t} |\varphi_{j,t}| \leq \Lambda$$

What this does:

- Extracts only systematic between-group differences
- Shrinks noise-driven differences to zero
- Λ chosen by AIC

LASSO Regularization

Problem: 800,000+ product categories $\rightarrow$ overfitting risk

Solution: Penalize income group differences

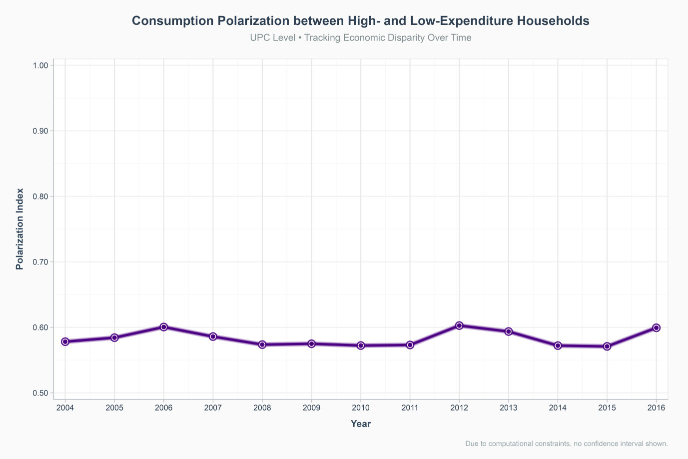
$$\sum_{j,t} |\varphi_{j,t}| \leq \Lambda$$

What this does:

- Extracts only systematic between-group differences
- Shrinks noise-driven differences to zero
- Λ chosen by AIC

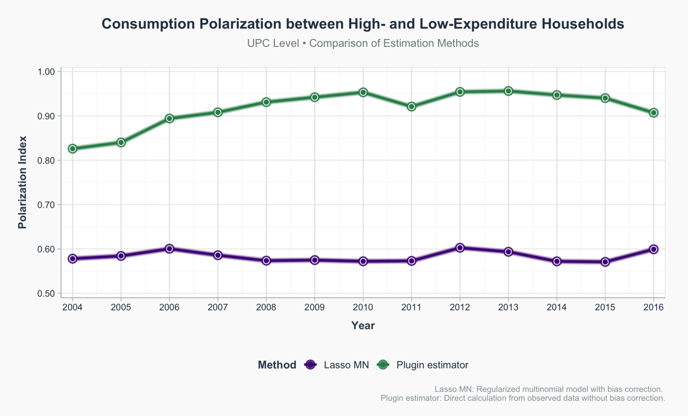
Implementation: Distributed Poisson approximation (Taddy, AoAS 2013)

Consumption Polarization: Predictive Power of UPC Codes



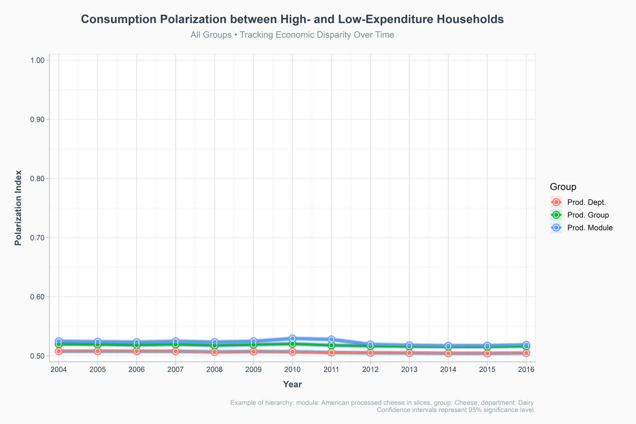
▸ Do we really need a model for choice probabilities?

Consumption Polarization: Predictive Power of UPC Codes



- **Regularized multinomial distribution:** captures systematic income differences, filters out noise
- **Plugin estimator:** upward bias due to data sparsity

Consumption Polarization: Predictive Power of Broader Categories



- ◇ With our methodology, broader categories always result in lower polarization values
- ◇ UPC level provides the **upper bound** for consumption polarization

▸ Where are Confidence Intervals?

Our Results vs. The Literature: A Contradiction?

Several recent papers using **similar data** reach conclusions that appear to conflict with ours:

- **Nord** (REStud R&R, 2024)
  Reports that the **overlap of purchases** between high- and low-income households is relatively low, with many products **predominantly purchased by only one group**. ▸ Histogram overlap vs. data sparsity

- **Bertrand & Kamenica** (AEJ: Applied, 2023)
  Using ensemble machine-learning methods, they show that certain products have **very high predictive power** for classifying a buyer's income group. ▸ Comparing Predictability-Based Measures

Our Results vs. The Literature: A Contradiction?

Several recent papers using **similar data** reach conclusions that appear to conflict with ours:

- **Nord** (REStud R&R, 2024)
  Reports that the **overlap of purchases** between high- and low-income households is relatively low, with many products **predominantly purchased by only one group**. Histogram overlap vs. data sparsity

- **Bertrand & Kamenica** (AEJ: Applied, 2023)
  Using ensemble machine-learning methods, they show that certain products have **very high predictive power** for classifying a buyer's income group. Comparing Predictability-Based Measures

The Puzzle

How do we reconcile these findings of **low overlap** and **high predictive power** at the product level with our result that our measure of **consumption polarization is very low**?

Main Questions

Research Questions ?

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?
3. What are implications for structural models of the answers to Questions 1 and 2?

Persistence in Basket Composition

Defining the Persistence Measure

Core Concept: We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

Defining the Persistence Measure

Core Concept: We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

$$O_i^E = \frac{\sum_{j \in (\mathcal{U}_{i,t-1} \cap \mathcal{U}_{i,t})} e_{i,t}(j)}{\sum_{j \in \mathcal{U}_{i,t}} e_{i,t}(j)}$$

$\mathcal{U}_{i,t-1}$ = products consumed by household i in year $t-1$

$\mathcal{U}_{i,t}$ = products consumed by household i in year t

$e_{i,t}(j)$ = expenditures on product j by household i in year t

Defining the Persistence Measure

Core Concept: We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

$$O_i^E = \frac{\sum_{j \in (\mathcal{U}_{i,t-1} \cap \mathcal{U}_{i,t})} e_{i,t}(j)}{\sum_{j \in \mathcal{U}_{i,t}} e_{i,t}(j)}$$

$\mathcal{U}_{i,t-1}$ = products consumed by household i in year $t-1$

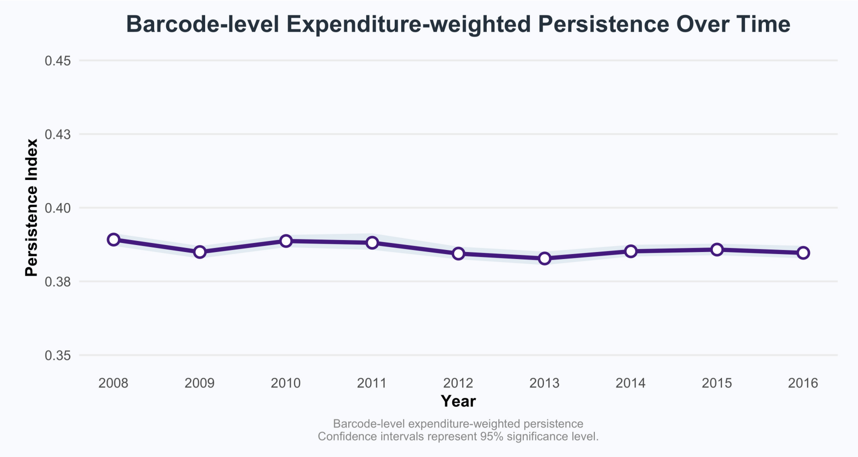
$\mathcal{U}_{i,t}$ = products consumed by household i in year t

$e_{i,t}(j)$ = expenditures on product j by household i in year t

Robustness checks:

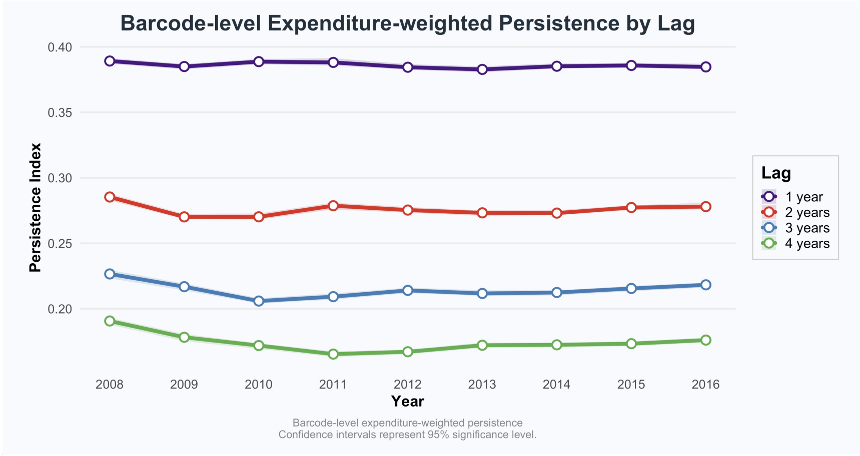
- Restrict analysis to products available in both years to control for market entry/exit effects
- Use higher-order lags
- Use different product definitions: ▪ UPC ▪ modules ▪ modules$\times$brands

Individual Consumption Baskets: Constantly in Flux



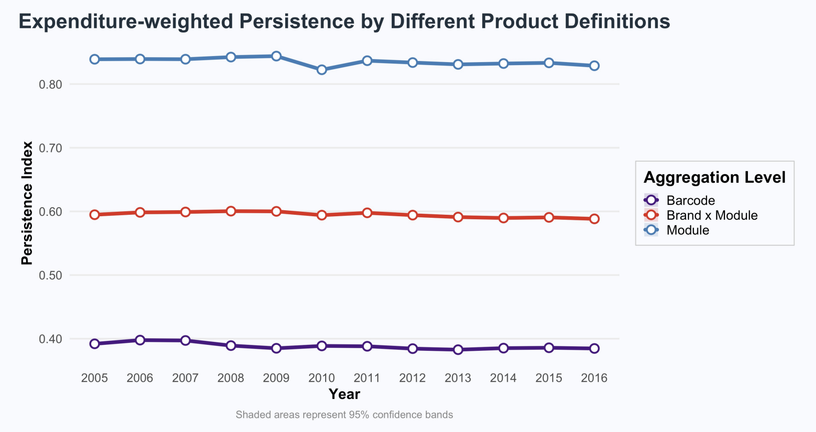
Key Insight: Household consumption patterns exhibit significant year-to-year variation, highlighting the dynamic nature of consumer choice behavior

Persistence Decreases with Longer Time Horizons



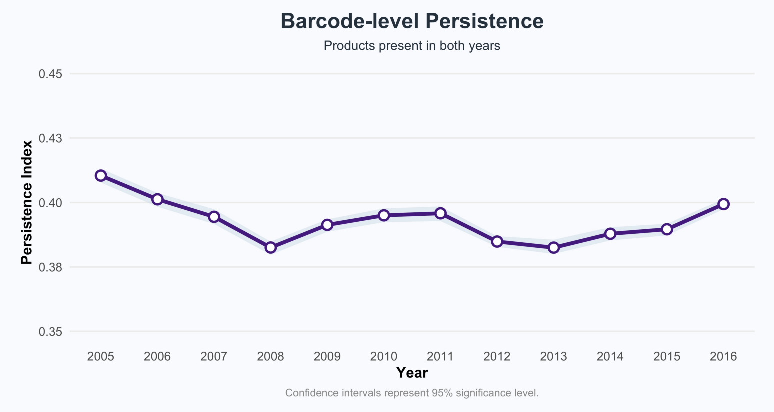
Pattern: Consumption persistence drops from 40% (1-year lag) to below 20% (4-year lag)

Product Definition Matters: Modules vs. Specific Products



Key Finding: Persistence jumps dramatically to over 80% when using product modules instead of specific UPC codes, revealing that households have stable *category* preferences but flexible *brand* choices

Robustness Check: Product Exit Does Not Drive Results



Robustness Check: Restricting analysis to products present in both years yields similar persistence patterns, confirming that our baseline findings are not driven by product entry or exit dynamics

Interpreting the Persistence Pattern

What This Tells Us About Consumer Behavior

Stable Category Preferences:

- Households know what general products they want (e.g., "American processed cheese in slices")
- These broader consumption patterns persist over time

Flexible Brand/Product Choices:

- Limited attachment to specific brands or UPC codes
- Willingness to substitute within product categories

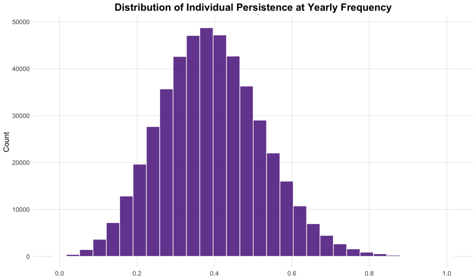
Implication: Consumers exhibit *categorical loyalty* rather than *brand loyalty*, maintaining consistent consumption needs while remaining flexible on specific product choices

Heterogeneity in Persistence

While average persistence appears quite low, significant **heterogeneity** exists across households.

Heterogeneity in Persistence

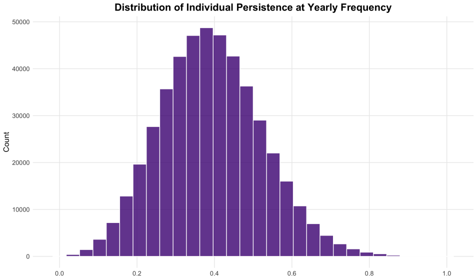
While average persistence appears quite low, significant **heterogeneity** exists across households.



- **Substantial variation:** Over 50% of aggregate non-durable consumption comes from households with persistence below 50%

Heterogeneity in Persistence

While average persistence appears quite low, significant **heterogeneity** exists across households.



- **Substantial variation:** Over 50% of aggregate non-durable consumption comes from households with persistence below 50%
- **Latent heterogeneity:** This variation in persistence is not explained by observable household characteristics

Main Questions

Research Questions ?

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?
3. What are implications for structural models of the answers to Questions 1 and 2?

Shopping Spree of Beckerian Consumers

Alternative View: Consumption as Random Sampling

- **Recent models:** Differences in consumption baskets result from idiosyncratic substitution patterns and heterogeneous preferences

Alternative View: Consumption as Random Sampling

- **Recent models:** Differences in consumption baskets result from idiosyncratic substitution patterns and heterogeneous preferences
- **Our extremely different thought experiment:** What if basket differences simply reflect random sampling because all products are (price-adjusted) perfect substitutes?
  Data-driven version of Becker's (JPE, 1962) impulsive consumers
  or "balls and bins" (Armenter & Koren, AER 2014) applied to consumption choices

Alternative View: Consumption as Random Sampling

- **Recent models:** Differences in consumption baskets result from idiosyncratic substitution patterns and heterogeneous preferences
- **Our extremely different thought experiment:** What if basket differences simply reflect random sampling because all products are (price-adjusted) perfect substitutes?

  Data-driven version of Becker's (JPE, 1962) impulsive consumers

  or "balls and bins" (Armenter & Koren, AER 2014) applied to consumption choices

Extreme example:

I Gave People $1,000,000 But ONLY 1 Minute To Spend It!

209M views • 3 years ago

MrBeast

(its epic) New Merch - https://**mrbeast.store** Check out Viewstats! - https://www.viewstat

5:16 Je veux remercier M. Jimmy, **Mr Beast** et tout le monde. Ça rendra la vie bien plus

CC

Shopping Spree Model

- ◇ Each household i has its own budget $m_{i,t}$ and nothing can be saved for the future.
- ◇ No information about products: to buy a product, they need to find it.
- ◇ There is no order in preferences; households randomly draw products during each purchase.
- ◇ Goal of each household: to spend as much as possible.

[1] This reflects our empirical finding of no consumption polarization between rich and poor households.
[2] $\mathbf{e}_j$ is a vector with 1 at the j-th position and 0 elsewhere.

Shopping Spree Model

- Each household i has its own budget $m_{i,t}$ and nothing can be saved for the future.
- No information about products: to buy a product, they need to find it.
- There is no order in preferences; households randomly draw products during each purchase.
- Goal of each household: to spend as much as possible.

Decision Problem of the Consumer

0. The consumer's consumption $\mathbf{c}_{i,t}$ is a zero vector at the beginning of the period.
1. The consumer randomly draws product j from the set of all products. The probability of drawing product j is product-specific and common across all consumers.[1] The number of purchased products, n_j, is drawn from a product-specific Poisson distribution.
2. Update $\mathbf{c}_{i,t} \leftarrow \mathbf{c}_{i,t} + \mathbf{e}_j n_j$.[2]
3. If the budget constraint for the new bundle is not violated ($\mathbf{p}'_{i,t}\mathbf{c}_{i,t} < m_{i,t}$), go to **Step 1**. Otherwise, **Stop**.

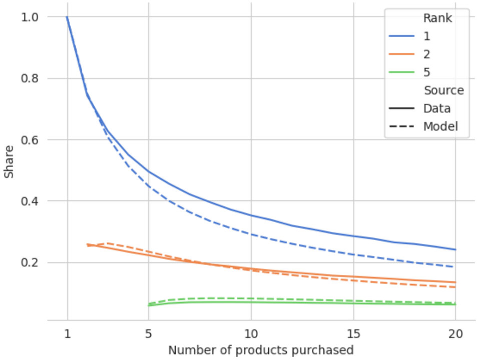
[1] This reflects our empirical finding of no consumption polarization between rich and poor households.
[2] $\mathbf{e}_j$ is a vector with 1 at the j-th position and 0 elsewhere.

Model Validation: Spending Concentration Plot

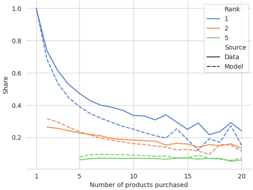
- **Diagnostic introduced by Neiman & Vavra (AEJ:Macro, 2023):**
  Households concentrate spending on top-ranked products, but concentration varies with consumption diversity
- **Plot structure:**
  - **X-axis:** Total products consumed ($|\Omega_i|$)
  - **Y-axis:** Average spending share on ranked products
  - **Lines:** Different ranks (1st, 2nd, 5th)
- **Validation approach:**
  - $\rightarrow$ **Solid** $=$ data patterns
  - $\rightarrow$ **Dashed** $=$ model predictions

Model Validation: Spending Concentration Plot

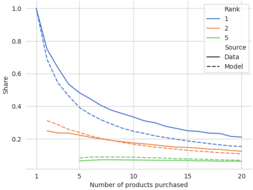
- **Diagnostic introduced by Neiman & Vavra (AEJ:Macro, 2023):** Households concentrate spending on top-ranked products, but concentration varies with consumption diversity
- **Plot structure:**
  - **X-axis:** Total products consumed ($|\Omega_i|$)
  - **Y-axis:** Average spending share on ranked products
  - **Lines:** Different ranks (1st, 2nd, 5th)
- **Validation approach:**
  - → **Solid** = data patterns
  - → **Dashed** = model predictions



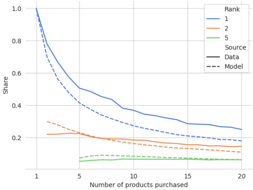
Spending Concentration within Different Product Categories



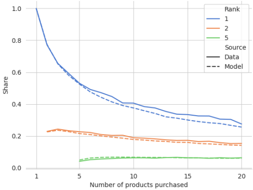
(a) Seafood Canned



(b) Cereal



(c) Yogurt



(d) Pet Food

Shopping Spree Model: Results

✔ **Success:** Model captures key cross-sectional features despite simplicity

✘ **Limitation:** Ad hoc dynamics generate far too little persistence

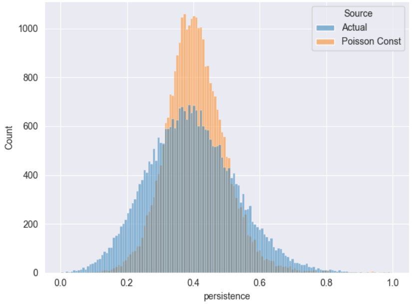
Need richer dynamic structure to match observed persistence

Shopping Spree Model + Consumer Inertia

- Introducing modest persistence, similar to Becker's (JPE, 1962) "inertia" concept, where each previously purchased product is repurchased with probability $\delta = 0.4$, yields compelling results.

Shopping Spree Model + Consumer Inertia

- Introducing modest persistence, similar to Becker's (JPE, 1962) "inertia" concept, where each previously purchased product is repurchased with probability $\delta = 0.4$, yields compelling results.
- Model replicates persistence **heterogeneity** despite calibration to *average dynamics only*
- Extreme values driven by households with **fewer transactions** Just like in data



Ok. . . so what?

Observational Equivalence Problem

Empirical patterns typically attributed to heterogeneous preferences can be generated by very different models

Ok. . . so what?

Observational Equivalence Problem

Empirical patterns typically attributed to heterogeneous preferences can be generated by very different models

Other examples of "stirring up a hornet's nest":

- "Balls and bins" vs. gravity models (Armenter & Koren, AER 2014)
- Price search vs. Dixit-Stiglitz (Menzio, JPE R&R 2025)
- Irrational vs. rational consumers (Becker, JPE 1962)
- Assortative wage matching vs. AKM (Borovičková & Shimer, QJE R&R 2024)

Our contribution:

Consumption differences from sampling histories alone—no preference heterogeneity needed

Why This Matters

Cautionary Tale

If very different models can generate similar empirical patterns, we must be cautious before attributing consumption differences to preferences.

Diverging Policy Implications: Consider the effect of increasing product variety:

- **Shopping Spree Model:** Welfare *decreases*
- **Heterogeneous Preferences:** Welfare *increases* (Neiman & Vavra, AEJ: Macro 2023)

Model identification is crucial for sound policy design.

Concluding Thoughts

1. **Consumption Similarity Across Income Groups**
   - Rich and poor households make surprisingly similar consumption choices
   - Product purchase data alone cannot distinguish household income groups

   Challenge: Intratemporal non-homothetic preference models

2. **Consumption Instability**
   - Individual choices are highly unstable over time
   - Repeat purchase probability: only 40% year-over-year

   Challenge: Latent heterogeneous preference models

3. **Beckerian Impulsive Consumers**
   - Randomness-driven model explains consumption basket differences
   - (Surprisingly) strong empirical fit with scanner data

Key Takeaway

The paper strikes a cautious note on policy implications drawn from models with heterogeneous preferences.

Thank you for your attention!

Group Differences in High-dimensional Choices

- **Context:** Gentzkow, Shapiro, and Taddy (Ecta, 2019) identify bias in their study on the polarization of US politics using congressional speech data

- **Model:** They developed a formal model of the speech-generating process

- **Solution:** They analyzed the bias formally and introduced estimators to overcome finite-sample bias

- **Result:** Their methods help recover a more accurate estimate of polarization

- **Application:** Their methodology is well-suited for studying consumption polarization due to structural similarities between speech and scanner data Gentzkow on Gentzkow et al. (Ecta, 2019)

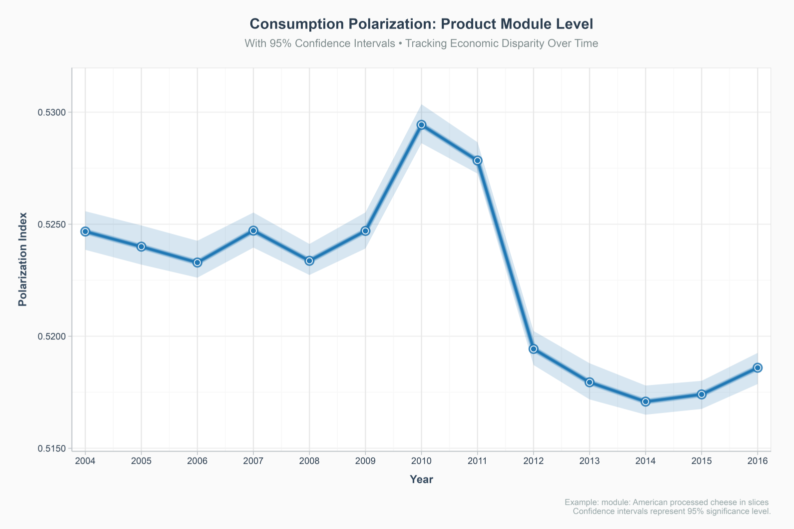
Back

Gentzkow on Gentzkow et al. (Ecta, 2019)

"Let's look at how different the choices of two different groups are in that high-dimensional space (. . .) It could be which UPCs did they buy in the supermarket."

Back

Consumption Polarization: Predictive Power of Modules



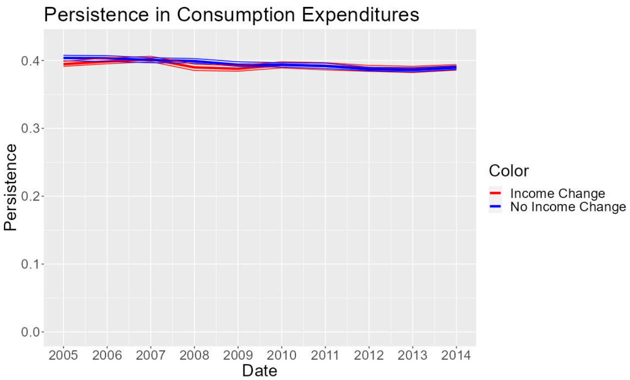
◂ Back to All Groups Overview

Group Averages

Quantile	Size	Male Age	Female Age	Race	Income
1st	2.16	4.38	6.24	1.38	20
2nd	2.26	5.22	6.54	1.29	20.2
3rd	2.32	5.65	6.63	1.25	20.1
4th	2.33	5.94	6.65	1.21	20
5th	2.19	6.16	6.37	1.17	19.3

Back

Persistence and Income Change



Back

Zooming into the Heterogeneity

Question

Can we attribute the heterogeneity to observable characteristics of the households?

Zooming into the Heterogeneity

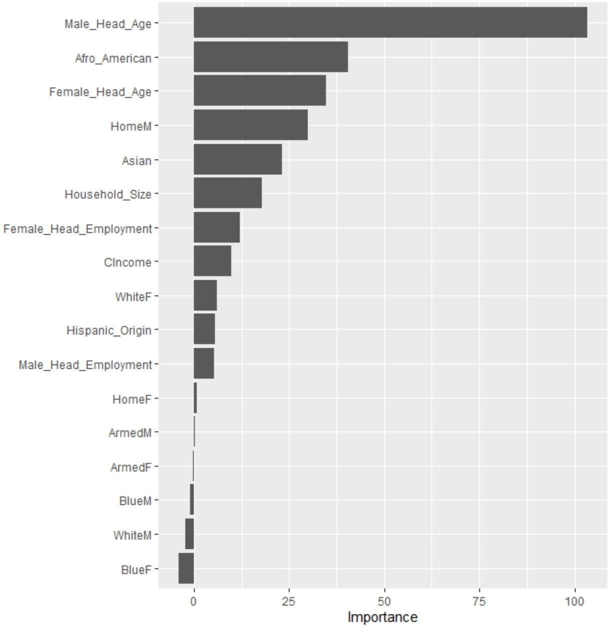
Question

Can we attribute the heterogeneity to observable characteristics of the households?

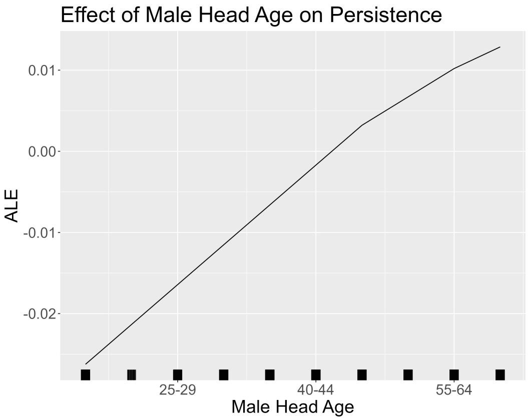
- We employ a random forest algorithm to find the household characteristics that are most important to explain differences within persistence
- We can then use accumulated local effects to see the size and direction of the influence of individual characteristics
  - Identifies the effect of singular variables while holding all else constant

Alternatively: Subgroup analysis

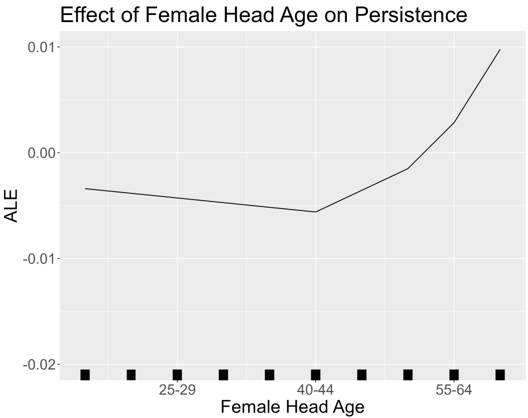
Variable Importance



Effect Age

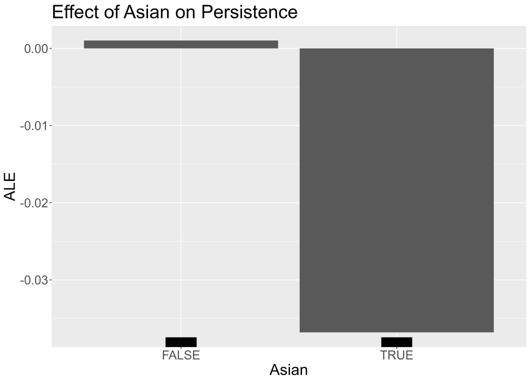


Effect of Male Head Age on Persistence
0.01
0.00
-0.01
-0.02
ALE
25-29
40-44
55-64
Male Head Age

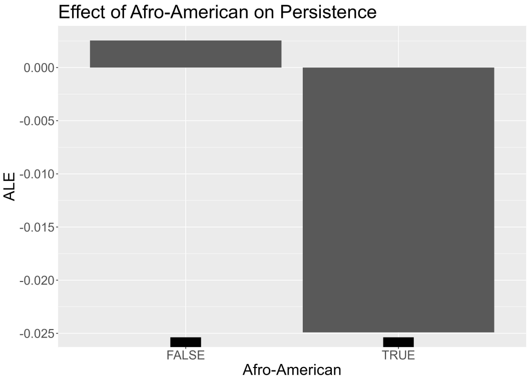


Effect of Female Head Age on Persistence
0.01
0.00
-0.01
-0.02
ALE
25-29
40-44
55-64
Female Head Age

Effect Race



Effect of Asian on Persistence
0.00
-0.01
-0.02
-0.03
ALE
FALSE
TRUE
Asian



Effect of Afro-American on Persistence
0.000
-0.005
-0.010
-0.015
-0.020
-0.025
ALE
FALSE
TRUE
Afro-American

Results: Persistence Summary

- We have seen that households are **on average** highly impersistent with their consumption choices
- More than 50% of aggregate non-durable consumption is accounted for by households with a persistence of less than 50%
- **But** there is substantial heterogeneity within persistence
- Most of the heterogeneity is latent and only some can be explained by household characteristics.

Back

Data Sparsity in Nielsen Dataset

Table 1: Fraction of Products by Minimum Transaction Threshold

		Minimum No. of Transactions			
Product Definition	**Geographic Scope**	≥ 1	≥ 2	≥ 10	≥ 20
UPC	Nationwide	1.000	0.963	0.814	0.723
UPC	Scantrack Market	1.000	0.713	0.287	0.209

- **Nationwide:** Even with national aggregation, 28% of products have fewer than 20 transactions
- **Local markets:** 79% of products have fewer than 20 transactions in local markets
- Sparsity increases dramatically when moving from national to local market analysis

Return to Main Slide

Some Products Are More Popular in Certain Income Groups

Soft'n Gentle Tissue

Unscented, 2-Ply

Purchased **1.98× more frequently** by **bottom expenditure quintile**

(vs. top quintile, 2008)

Tropicana Pure Premium

Orange Juice, 128 Oz

Purchased **7.56× more frequently** by **top expenditure quintile**

(vs. bottom quintile, 2008)

Go Back

Universal Appeal Across Income Distribution

Kellogg's All-Bran

Original, 18.5 Oz

Purchased at **similar frequencies** by both **high and low income** consumers

(Less than one-percent difference between quintiles, 2008)

Go Back

Consumption Polarization: Raw Data Approach

Plugin Estimator: Direct calculation from raw data. **Spending probability** for income group P can be computed directly:

$$\hat{q}^{P}_{j,t} = \frac{\sum_{i \in P} \sum_j c_{ijt}}{\sum_{i \in P} m_{it}}$$

- c_{ijt}**:** household i consumption on product j
- m_{it}**:** total expenditure by household i
- P**:** income group (high/low expenditure)

Consumption Polarization: Raw Data Approach

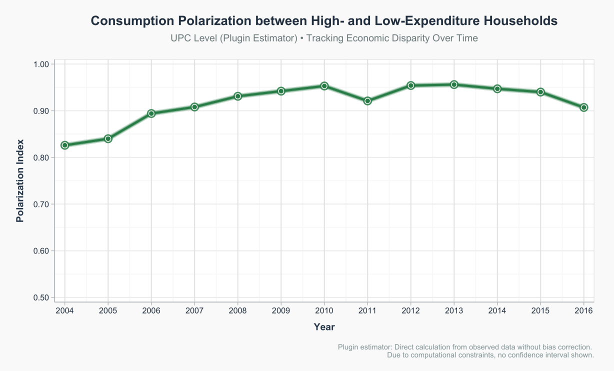
Plugin Estimator: Direct calculation from raw data. **Spending probability** for income group P can be computed directly:

$$\hat{q}^{P}_{j,t} = \frac{\sum_{i \in P} \sum_{j} c_{ijt}}{\sum_{i \in P} m_{it}}$$

c_{ijt}: household i consumption on product j

m_{it}: total expenditure by household i

P: income group (high/low expenditure)



Bias alert

Without correction for data sparsity, polarization $\bar{\pi}_t$ is artificially overestimated.
The plugin estimator inflates disparity levels compared to our model-based approach.

Back

Basket Similarity: Setup & Previous Findings

Histogram overlap due to Nord (REStud R&R, 2024):

$$\boxed{\Omega^{g,h} = \sum_j \min\{\omega_j^g, \omega_j^h\}}$$

where:

$$\omega_j^g = \frac{\sum_{i \in g} e_j^i}{\sum_k \sum_{i \in g} e_k^i}.$$

$\Omega^{H,L} = 1:$ Similar basket composition (homothetic preferences)

$\Omega^{H,L} < 1:$ Different basket composition (non-homothetic preferences)

Basket Similarity: Setup & Previous Findings

Histogram overlap due to Nord (REStud R&R, 2024):

$$\boxed{\Omega^{g,h} = \sum_j \min\{\omega_j^g, \omega_j^h\}}$$

where:

$$\omega_j^g = \frac{\sum_{i \in g} e_j^i}{\sum_k \sum_{i \in g} e_k^i}.$$

$\Omega^{H,L} = 1$: Similar basket composition (homothetic preferences)

$\Omega^{H,L} < 1$: Different basket composition (non-homothetic preferences)

Strong polarization between rich and poor households at the barcode level:

✓ **Modules:** $\Omega^{H,L} \approx 0.86$

× **Barcodes:** $\Omega^{H,L} \approx 0.63$

Statistical Diagnostics: Dissecting Histogram Overlap

Problem

- Many products bought rarely
- Sampling noise $\Rightarrow \min\{\omega_j^H, \omega_j^L\}$ small
- Bias persists **even with identical preferences**

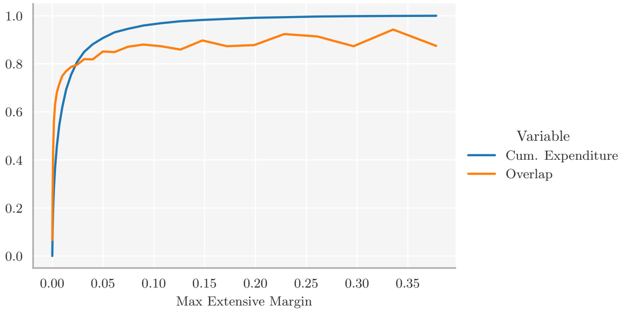
Statistical Diagnostics: Dissecting Histogram Overlap

Problem

- Many products bought rarely
- Sampling noise $\Rightarrow \min\{\omega_j^H, \omega_j^L\}$ small
- Bias persists **even with identical preferences**

Evidence

- Products bought by $< 1\%$: very low overlap
- Products bought by $\geq 1\%$: overlap > 0.8
- Rare products $= 80\%$ of spending



Implication:

Histogram overlap may suffer from **small sample bias**

Placebo Test: Quantifying the Bias

Experimental Design

1. Randomly assign households to income quintiles
2. Compute overlap between "rich" and "poor" quintiles
3. Each quintile should have the **same distribution** of individuals & choices

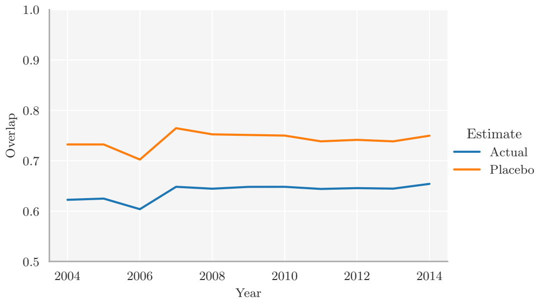
Placebo Test: Quantifying the Bias

Experimental Design

1. Randomly assign households to income quintiles
2. Compute overlap between "rich" and "poor" quintiles
3. Each quintile should have the **same distribution** of individuals & choices

Expected vs. Actual

- **Expected:** $\Omega^{\text{Placebo}} = 1$ (no bias)
- **Actual:** $\Omega^{\text{Placebo}} \ll 1$ (bias detected!)
- Deviations from 1 $\Rightarrow$ histogram overlap is biased



Key Finding:

Up to **3/4** of observed drop is **statistical bias**

Back

Comparing Predictability-Based Measures

- Bertrand & Kamenica (AEJ: Applied, 2023) find high predictive power of consumption behavior for income classification ($\approx 90\%$)
- Yet our results are much more modest ($< 60\%$, where 50% is the lower bound)

Comparing Predictability-Based Measures

- Bertrand & Kamenica (AEJ: Applied, 2023) find high predictive power of consumption behavior for income classification ($\approx 90\%$)
- Yet our results are much more modest ($< 60\%$, where 50% is the lower bound)

How come?

Comparing Predictability-Based Measures

- Bertrand & Kamenica (AEJ: Applied, 2023) find high predictive power of consumption behavior for income classification ($\approx 90\%$)
- Yet our results are much more modest ($< 60\%$, where 50% is the lower bound)

How come?

Cultural Distance
Bertrand & Kamenica (AEJ:Applied, 2023)

- **Focus:** *Existence* of distinguishing behaviors
- **Method:** Maximum classification accuracy across all possible product combinations

Consumption Polarization
(Our Approach)

- **Focus:** *Magnitude* of distinguishing behaviors in overall spending
- **Method:** Expected information value of the "representative" dollar spent

Key Distinction

Bertrand & Kamenica: Identifies *whether* household groups have distinct consumption behaviors

Our approach: Quantifies *how much* goods that distinguish households contribute to overall grocery spending

Illustrative Example: Dinner at Beppa Fioraia

Consumption Breakdown

Item	Me	Pavel
Bistecca Fiorentina	30€	30€
Chianti Classico	10€	10€
Tiramisu	7€	7€
Spinacci	**3€**	–
French Fries	–	**3€**
Total	**50€**	**50€**

Illustrative Example: Dinner at Beppa Fioraia

Consumption Breakdown

Item	Me	Pavel
Bistecca Fiorentina	30€	30€
Chianti Classico	10€	10€
Tiramisu	7€	7€
Spinacci	**3€**	–
French Fries	–	**3€**
Total	**50€**	**50€**

Methodological Results

Bertrand & Kamenica (AEJ:Applied, 2023)

- Side dishes allow **100%** classification accuracy
- Consequently: Cultural Distance = **Maximum**

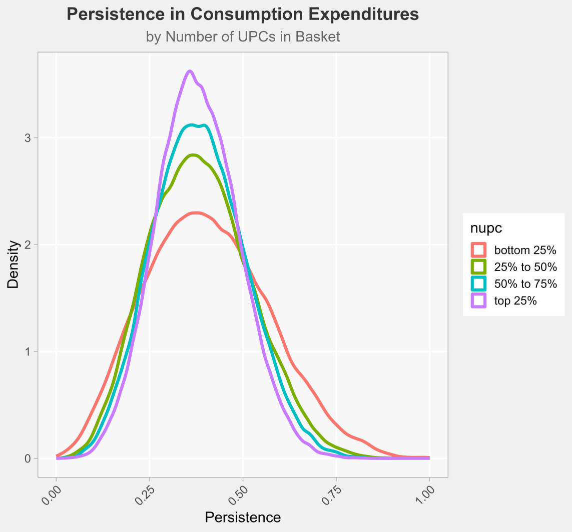
Our Consumption Polarization:

$$\text{Polarization} = \frac{2 \times 3 \times 100\% + 2 \times 47 \times 50\%}{100}$$

$$= \mathbf{53\%} \ \text{(Close to lower bound, 50\%)}$$

Back

Persistence and Basket Composition



Households with larger baskets (more UPCs) exhibit more concentrated persistence distributions around the average value.

Back