Mostly rational types

Andrew Tai*

July 2026
[Click here for most recent draft.]

Abstract

This paper proposes a method combining revealed preference theory and machine learning ideas to partition consumers in a principled way. The Generalized Axiom of Revealed Preferences can be applied to partition consumers into fully rational types, in the classical consumer theory sense. From here, types can be combined in a minimally lossy way. I illustrate using household milk purchase data. In the example data, 15 types are required to rationalize the data; however, many of these have few households, and can be merged into other types without significantly harming rationality. My method thus establishes a theory-driven and principled way to describe consumer demand data and the tradeoff between parsimony and heterogeneity.

1 Introduction

Gary Becker famously advocated for an approach to economics embracing assumptions of maximizing behavior, stable preferences, and equilibrium:

> The assumption of stable preferences provides a stable foundation for generating predictions about responses to various changes, and prevents the analyst from succumbing to the temptation of simply postulating the required shift in preferences to 'explain' all apparent contradictions to his predictions. Becker, 1976.

Simply, if economics is a science, then unexplained phenomena must not be consistently hand-waved away as unique phenomena. Stigler and Becker (1977) are even more stark:

> Tastes neither change capriciously nor differ importantly between people.

However, this poses a problem for the estimation of consumer behavior, where significant individual heterogeneity is natural. While some of this heterogeneity can be modeled on the consumers' observables (e.g. wealth, demographics, etc.), the majority is an accident of nature.

*US Dept. of War. Contact: andrew.a.tai@gmail.com. Disclaimer: The views expressed are those of the author and do not reflect the official policy or position of the Department of War or the U.S. Government.

Given consumption data from a panel of households, the classic approach is to pool the data, and control on household observables if available. An early example, Banks, Blundell, and Lewbell (1997), finds that the approach works well for some categories of purchases but very poorly for others. They apply the (quadratic) almost ideal demand system, which explicitly assumes a representative consumer. More recent models like Berry, Levinsohn, and Pakes (2004) allow random coefficients to incorporate individual heterogeneity (among other desiderata). Nevertheless, when households have significant unobservable heterogeneity, the resulting estimates may be imprecise; and point estimates may not accurately describe any households at all.

Crawford and Pendakur (2013) propose to apply revealed preferences, chiefly Afriat (1967) and Varian (1982), to partition households into fully rational types. Afriat's theorem establishes the Generalized Axiom of Revealed Preference (GARP) as a necessary and sufficient condition for a set of consumption data to be generated by a well-behaved utility function $u(x)$. Given an observed choice x_i at a price p_i, it must dominate any other affordable choice. The following is an example of a violation:

- Bundle x_1 was chosen at price p_1
- Bundle x_2 was chosen at price p_2
- $p_1x_1 \geq p_1x_2$ – that is, x_2 was affordable when x_1 was chosen
- $p_2x_2 \geq p_2x_1$ – that is, x_1 was affordable when x_2 was chosen
- At least one of the inequalities is strict

This violation implies a cycle in preferences, $x_1 \succeq x_2$ and $x_2 \succeq x_1$, with at least one strict. GARP generalizes this idea from 2-cycles to general length.

GARP is usually interpreted as a check on repeated observations from one individual (household). Crawford and Pendakur propose to use GARP to partition household data into rational types. That is, they ask *how many distinct utility functions are generating the consumer demand data?*

I apply this method to data from the USDA, National Household Food Acquisition and Purchase Survey (FoodAPS). I find that 15 types are necessary to rationalize the 2,167 households. However, many of these types contain very few households. I thus propose to combine this method with ideas in machine learning for classification like hierarchical clustering and KNN. Many groups can be combined while incurring low "loss" to rationalizability, calculating loss as either Critical Cost Efficiency Index (CCEI) or the Varian index (cf. Afriat (1973) and Varian (1990)). Using Varian's rule-of-thumb of 95% rationality as essentially rational, we can reduce the number of types by half. My method thus offers intermediate options between two extremes – assuming all households have the same underlying utility; or requiring full rationality. By sacrificing a modest amount of rationality (interpretable as requiring individual heterogeneity), we require far fewer types.

The next section describes the dataset. Section 3 describes the proposed method to partition the households into groups. Section 4 estimates demand via almost ideal demand system (AIDS). Section 5 concludes.

2 Data

To illustrate my proposed method, I use data from the USDA FoodAPS survey, which was fielded from April 2012 through mid-January 2013. Each participating household reported the foods that household members acquired over a 7-day period. The final analytic sample was 4,826 households. I focus particularly on household milk purchases, partitioned into four types of milk: whole, 2%, 1%, and skim/nonfat milk. Among the survey sample, 2,167 households are observed purchasing any milk; this forms my sample.

For each household, I observe the quantity of each type of milk purchased and the prices at which they were purchased. Prices were generally not recorded for goods not purchased; thus I impute prices when not recorded by the median in the matching category: region, rural / urban status, and store type. While FoodAPS level of observations is a purchase event (e.g. one receipt), I aggregate each household to a weekly sum. Table 1 summarizes the data. Note the extreme outlier shown in the maximum price of whole milk. I chose to leave this observation in the dataset, as the objective of this paper is to illustrate my proposed method, rather than to accurately estimate milk demand.

Table 1: Descriptive statistics ($N = 2167$)

	Mean	Min	Max	SD
Budget shares				
Whole	0.29	0.00	1.00	0.43
2%	0.46	0.00	1.00	0.47
1%	0.10	0.00	1.00	0.29
Skim	0.13	0.00	1.00	0.33
Total expenditure ($)				
Total expenditure	4.97	0.30	88.47	4.30
Prices ($ per 100 g)				
Whole	0.14	0.02	14.36	0.53
2%	0.10	0.02	0.79	0.06
1%	0.10	0.01	0.75	0.06
Skim	0.11	0.02	1.12	0.08

This paper does not seek to argue that estimating household milk demand is important in and of itself. Rather, this setting readily illustrates the issues that this paper seeks to note. While it is entirely possible to estimate a demand model on household milk purchases, the presence of

individual heterogeneity makes this difficult. There may be individual heterogeneity by taste for milk (e.g. whole vs. skim) or by price sensitivity (e.g. some households are more price sensitive, or more likely to substitute between milk types). Even with household characteristics, it may be difficult to describe this heterogeneity – much of it is likely due to unobservable taste.

3 Method

I now turn to a proposed method for partitioning households into distinct types. Based on the observed consumption data, the initial exercise prescribed by Crawford and Pendakur (2013) is to partition households into groups, each of which satisfies GARP. That is, each group can be treated as having a single utility function. This enables us to interpret the errors from any demand estimation model purely as model misspecification error rather than individual heterogeneity.

The core of this method is Afriat's theorem.

Theorem 1. *A set of consumer demand data* $\{(x_i, p_i)\}_{i \in N}$ *can be described by a maximization of a utility function if and only if it satisfies the generalized axiom of revealed preference (GARP).*

I refer the reader to Afriat (1967) for details (or any number of lecture notes available online; e.g. by Echenique). Crawford and Pendakur propose an infeasible algorithm, Algorithm 1, to partition the data into the fewest number of groups, each of which satisfies GARP.

Algorithm 1 Exact minimum-type partition

Require: Observations $I = \{1, \ldots, n\}$

Ensure: Minimum number of types $N^\star$ and a rationalizing partition $\mathcal{G}^\star$

```
for K = 1, ..., n do
    for each partition G = {G_1, ..., G_K} of I into K nonempty sets do
        if G_k satisfies GARP for every k = 1, ..., K then
            N* ← K
            G* ← G
            return (N*, G*)
        end if
    end for
end for
```

This algorithm is infeasible, since it requires checking GARP for every partition of the dataset. Thus it requires checking GARP a combinatorial number of times; even though GARP itself can be checked in $O(N^3)$ time.

Algorithm 2 calculates a feasible upper bound on the number of types. The observations are first randomly ordered. The first observation forms the first group. The next observation is checked for GARP consistency with the first group. If they are consistent, then it is added to this group;

otherwise it is added to a new group. Each subsequent observation undergoes the same procedure. The entire process can be repeated multiple times, and the best (smallest) partition is chosen.

Algorithm 2 Crawford–Pendakur upper-bound algorithm

Require: Observations $I = \{1, \ldots, n\}$ and number of restarts R
Ensure: Upper bound $\bar{N}$ and a rationalizing partition $\bar{\mathcal{G}}$

$\bar{N} \leftarrow \infty$
for $r = 1, \ldots, R$ **do**
  Draw a random ordering $(i_1, \ldots, i_n)$ of I
  $\mathcal{G} \leftarrow \varnothing$
  for $t = 1, \ldots, n$ **do**
    $\mathcal{C} \leftarrow \{G \in \mathcal{G} : G \cup \{i_t\} \text{ satisfies GARP}\}$
    if $\mathcal{C} \neq \varnothing$ **then**
      Choose $G^\star \in \arg\max_{G \in \mathcal{C}} |G|$
      Replace $G^\star$ by $G^\star \cup \{i_t\}$ in $\mathcal{G}$
    else
      $\mathcal{G} \leftarrow \mathcal{G} \cup \{\{i_t\}\}$
    end if
  end for
  if $|\mathcal{G}| < \bar{N}$ **then**
    $\bar{N} \leftarrow |\mathcal{G}|$
    $\bar{\mathcal{G}} \leftarrow \mathcal{G}$
  end if
end for
return $(\bar{N}, \bar{\mathcal{G}})$

Figure 1 shows Algorithm 2 applied to the FoodAPS data. Over half of households can be described by a single utility function; there are 5-10 sizable groups, and the remainder are very small. Table 2 shows summary statistics for milk purchases by group; the descriptive statistics already illustrate significant heterogeneity in behavior across types. For example, Type 1 favors whole milk, Type 2 favors 2% milk, and Type 3 almost exclusively purchases 1% milk.

Motivated by this pattern, we may ask how much we gain in terms of "rationality" by having these small groups. Stated another way, do we lose much by merging some groups into others? Doing comes with the practical advantage of allowing us to estimate fewer demand models; and the philosophical advantage of respecting Stigler's and Becker's charge of limiting capricious changes in taste.

Of course, when we combine groups, by definition they must become non-rationalizable. That is, treating the households in the combined group as coming from a single utility function means that some of them behave irrationally. A number of indexes exist to quantify the degree of irrationality.

Figure 1: Partition of FoodAPS households into fully rational types

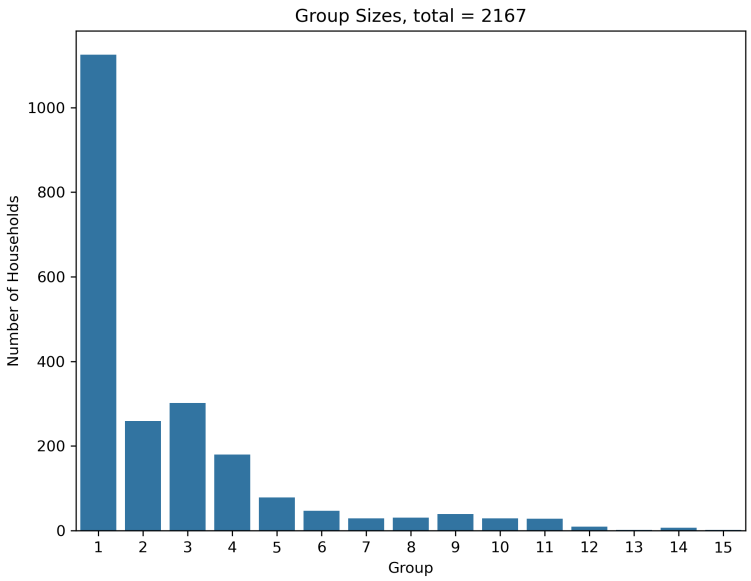


Table 2: Descriptive statistics by group

	N	Whole	2%	1%	Skim
Pooled	2167	0.29	0.47	0.10	0.14
Type 1	1125	0.23	0.57	0.09	0.12
Type 2	302	0.62	0.13	0.14	0.11
Type 3	259	0.03	0.95	0.02	0.00
Type 4	180	0.44	0.10	0.17	0.29
Type 5	78	0.40	0.07	0.11	0.42
Type 6	47	0.20	0.62	0.03	0.16
Type 7	39	0.88	0.07	0.02	0.03
Type 8	31	0.02	0.03	0.00	0.95
Type 9	29	0.59	0.03	0.34	0.04
Type 10	29	0.04	0.94	0.00	0.02
Type 11	28	0.01	0.11	0.68	0.19
Type 12	9	0.04	0.05	0.86	0.05
Type 13	7	0.84	0.04	0.00	0.12
Type 14	2	0.00	0.80	0.00	0.20
Type 15	2	0.00	0.00	0.73	0.27

In this paper, I use the critical cost efficiency index (CCEI) and the related Varian index. Intuitively, CCEI asks *how much would we need to shrink each observed budget before the consumer's choices become consistent with utility maximization?* A $CCEI = 1$ is completely rational; a $CCEI = 0.95$ is rational after allowing up to 5% of each household's budget to be wasted.

I now turn to my proposed algorithm for combining groups. We balance the desire for fewer and larger groups with the loss in rationality in each group. For a partition of households $\mathcal{P} = \{G_1, ..., G_k\}$, for a set of households G define the loss of treating all G as one type as

$$L(G) = |G|(1 - e_G)$$

where e_G is CCEI of G. For each pair of groups G_a, G_b, calculate the incremental merge cost as

$$L(G_a \cup G_b) - L(G_a) - L(G_b).$$

At each step, we merge the two groups with the lowest incremental merge cost. The final product is a curve between summed loss and the number of types, akin to hierarchical clustering or KNN. Greater loss is interpretable as requiring more individual heterogeneity to rationalize the data. Figure 2 shows the curve of loss and types. At 15 types, which fully rationalizes the data, there is 0 loss. Combining all types into one group results in significant loss (that is, significant individual heterogeneity). However, we can reduce the number of types to 8 while maintaining that rationality within type, on average. Informally, it is *approximately correct* to treat the households as sharing the same utility function.

I also repeat the analysis for the Varian index, which allows each household to waste an individual amount of its budget, and takes the average. The resulting measure is smoother and less punishing. Figure 3 shows the curve of Varian index loss to number of types.

4 Demand estimation

Incorporating individual heterogeneity into demand estimation is difficult; see for example Matzkin (2003). However, the almost ideal demand system (AIDS) is easy to estimate, at the cost of assuming away individual heterogeneity. The previous section's method offers a principled way to quantify the size of this assumption.

To illustrate, I estimates AIDS models by type. For household i and milk type j, the share equation is

$$w_i^j = \alpha_j + \sum_{k=1}^{4} \gamma_{jk} \ln p_i^k + \beta_j \ln\left(\frac{x_i}{P_i}\right) + \varepsilon_i^j$$

where

$$\ln P_i = \alpha_0 + \sum_k \alpha_k \ln p_i^k + \frac{1}{2} \sum_k \sum_\ell \gamma_{kl} \ln p_i^k \ln p_i^l$$

Figure 2: Loss - types curve for CCEI

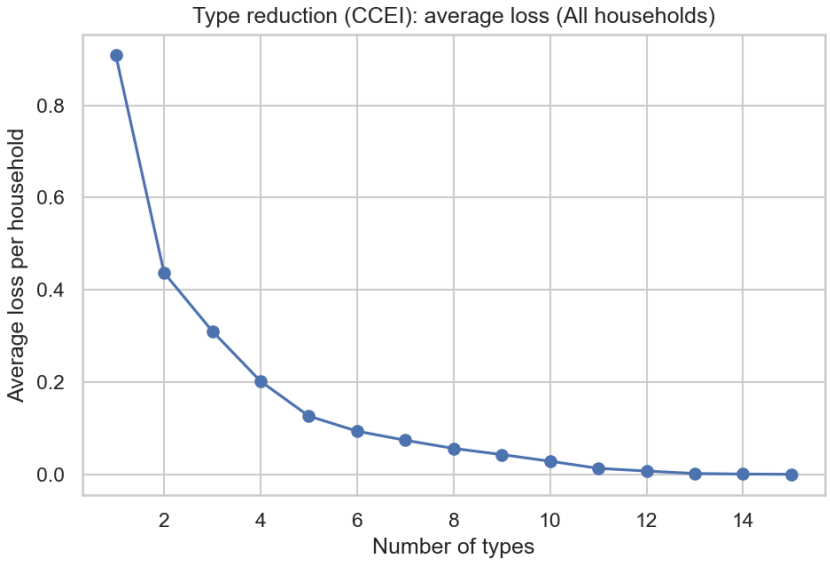
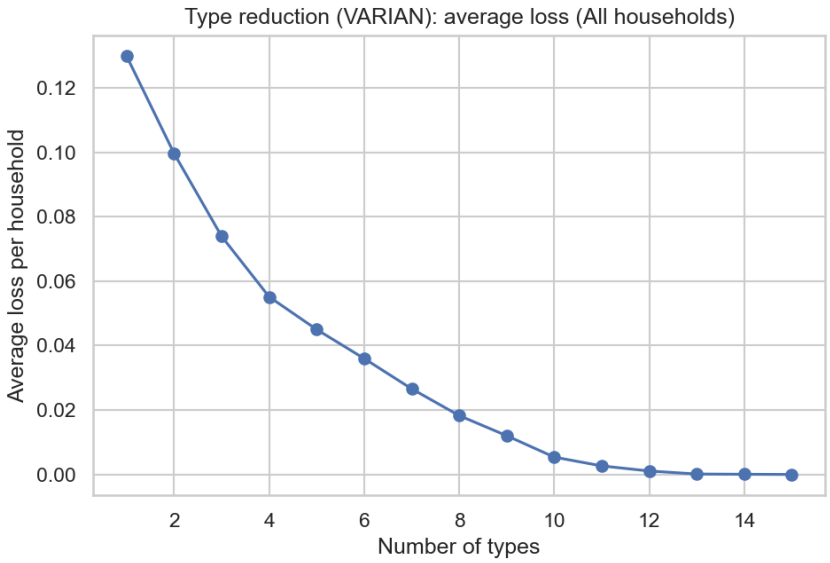


Figure 3: Loss - types curve for Varian index



and w_i^j is i's share of the milk budget spent on j, and p_i^j is the price of milk j faced by household i. Prices and expenditures are normalized so that $\ln p^j = \ln x = 0$ at the median household. Restrictions are

- $\sum_k \alpha_k = 1$ and $\sum_k \beta_k = 0$ – expenditure shares over milk categories sum to 1
- $\sum_k \gamma_{kl} = 0$ for all l – homogeneity of degree zero in prices and total expenditure
- $\gamma_{kl} = \gamma_{lk}$ – symmetry of the Slutsky matrix

The α_j are interpreted as predicted budget shares at the median prices and expenditure. β_j is expenditure semi-elasticity of the share, $\partial w_j / \partial \ln x$; that is, the change in j's share of the budget per log change in total milk expenditure. Intuitively, $\beta_j > 0$ implies milk j is a luxury, and $\beta_j < 0$ implies it is a necessity. γ_{kl} is the change in the share of k per unit change in the log of the price of l, $\ln p^l$, holding real expenditure fixed.

Table 3 shows AIDS α and β estimates for partition into 8 types (according to the CCEI procedure) as an example. The pooled sample estimates are incorrect on theoretical grounds (since we know we cannot treat all households as sharing a common utility function). Further, in practical terms, the estimates mask considerable heterogeneity. The pooled estimates are unsurprisingly close to the largest type, Type 1. But other large groups show very different behavior in both preferred milk and price responses. For example, Type 3 purchases almost exclusively 2% milk, but shows a modest negative expenditure semi-elasticity for it.

5 Conclusion

I propose a method combining revealed preference theory and machine learning ideas to partition consumers in a principled way. The Generalized Axiom of Revealed Preferences can be applied to partition consumers into fully rational types, in the classical consumer theory sense. From here, types can be combined in a minimally lossy way.

In the example data, 15 types are required to rationalize household milk purchases; however, many of these have few households, and can be merged into other types without significantly harming rationality. My method thus establishes a theory-driven and principled way to describe consumer demand data and the tradeoff between parsimony and heterogeneity. This matters in practice: application of AIDS models shows significant heterogeneity between household types.

Table 3: AIDS levels and expenditure semi-elasticities by CCEI-reduced type (all households, 8 types). Standard errors in parentheses; 0.000 (–) marks a milk type the group never purchases.

	N	Whole	2%	1%	Skim
Levels, a^j					
Pooled	2167	0.288	0.471	0.100	0.141
		(0.009)	(0.010)	(0.006)	(0.007)
Type 1	1125	0.249	0.518	0.101	0.132
		(0.012)	(0.014)	(0.008)	(0.009)
Type 2	302	0.640	0.072	0.163	0.125
		(0.027)	(0.017)	(0.019)	(0.018)
Type 3	259	-0.003	0.972	0.030	0.001
		(0.011)	(0.015)	(0.009)	(0.002)
Type 4	180	0.431	0.060	0.208	0.302
		(0.033)	(0.017)	(0.029)	(0.030)
Type 5	145	0.278	0.278	0.320	0.124
		(0.026)	(0.021)	(0.024)	(0.016)
Type 6	78	0.375	0.075	0.052	0.498
		(0.038)	(0.027)	(0.028)	(0.053)
Type 7	47	0.207	0.494	0.029	0.270
		(0.027)	(0.038)	(0.021)	(0.030)
Type 8	31	0.031	0.035	0.000	0.934
		(0.021)	(0.032)	–	(0.039)
Expenditure semi-elasticities, b^j					
Pooled	2167	-0.002	0.014	0.017	-0.029
		(0.015)	(0.016)	(0.010)	(0.011)
Type 1	1125	-0.011	0.016	0.032	-0.037
		(0.020)	(0.024)	(0.013)	(0.016)
Type 2	302	0.042	0.233	-0.236	-0.039
		(0.051)	(0.031)	(0.035)	(0.033)
Type 3	259	0.038	-0.050	0.008	0.004
		(0.010)	(0.014)	(0.009)	(0.002)
Type 4	180	-0.071	0.154	0.048	-0.131
		(0.037)	(0.019)	(0.031)	(0.033)
Type 5	145	0.051	-0.003	-0.040	-0.008
		(0.037)	(0.030)	(0.035)	(0.022)
Type 6	78	-0.348	0.094	0.281	-0.027
		(0.056)	(0.040)	(0.044)	(0.074)
Type 7	47	0.205	0.105	-0.012	-0.298
		(0.035)	(0.048)	(0.024)	(0.040)
Type 8	31	0.008	0.027	0.000	-0.036
		(0.020)	(0.034)	–	(0.039)

References

Afriat, S.N., 1967. The construction of utility functions from expenditure data. International economic review, 8(1), pp.67-77.

Afriat, S.N., 1973. On a system of inequalities in demand analysis: an extension of the classical method. International economic review, pp.460-472.

Banks, J., Blundell, R. and Lewbel, A., 1997. Quadratic Engel curves and consumer demand. Review of Economics and statistics, 79(4), pp.527-539.

Becker, G.S., 1976. The economic approach to human behavior (Vol. 803). University of Chicago press.

Berry, S., Levinsohn, J. and Pakes, A., 2004. Differentiated products demand systems from a combination of micro and macro data: The new car market. Journal of political Economy, 112(1), pp.68-105.

Crawford, I. and Pendakur, K., 2013. How many types are there?. The Economic Journal, 123(567), pp.77-95.

Deaton, A. and Muellbauer, J., 1980. An almost ideal demand system. The American Economic Review, 70(3), pp.312-326.

Matzkin, R.L., 2003. Nonparametric estimation of nonadditive random functions. Econometrica, 71(5), pp.1339-1375.

Stigler, G.J. and Becker, G.S., 1977. De gustibus non est disputandum. The american economic review, 67(2), pp.76-90.

Varian, H.R., 1982. The nonparametric approach to demand analysis. Econometrica: Journal of the Econometric Society, pp.945-973.

Varian, H.R., 1990. Goodness-of-fit in optimizing models. Journal of Econometrics, 46(1-2), pp.125-140.