

How Many Types Are There?

Ian Crawford and Krishna Pendakur

Oxford University and Simon Fraser University

"Do not do unto others as you expect they should do unto you. Their tastes may not be the same."

George Bernard Shaw, Man and Superman (Maxims for Revolutionaries) 1903

The Chicago School(s)

“Tastes neither change capriciously nor differ importantly between people.” Becker and Stigler De Gustibus Non Est Disputandum, AER, 1977

“Research in microeconometrics demonstrated that it was necessary to be careful in accounting for the sources of manifest differences among apparently similar individuals. ... This heterogeneity has profound consequences for economic theory and for econometric practice.” J. Heckman, Nobel Lecture, JPoE, 2001

The current consensus on unobserved heterogeneity in microdata is:

- 1 There's a lot of it (Lewbel and Pendakur, 2009).
- 2 Ignoring it won't do—you're often stuck with linearity (e.g., Lewbel 2001, Lewbel and Ng 2005).
- 3 It makes estimation/identification very tricky (Matzkin 2005, Hoderlein 2009).
- 4 It's hard to deal with (weird nonlinear quantile models, etc).
- 5 We ask: how much do you really need?

The Aim of Our Paper

- 1 This paper is about *preference* heterogeneity, but the model works for technology heterogeneity, too.
- 2 Cross-sectional data, but it works in panel settings, too.
- 3 We aim to find the *minimum* number of utility functions necessary to rationalize all observed behaviour.
- 4 n types will always work, but n isn't minimal—Occam's Razor (and Friedman 1953) tells us to seek the minimal model.
 - 1 if the minimum number of types is big (e.g., close to n), then fixed effects models, or models with a continuum of unobserved types may be good.
 - 2 if the minimum number of types is small, then discrete type models as in macro-labour, education choice or marketing might be good

Our framework offers three benefits to complement standard approaches to unobserved heterogeneity

- ① It is model-driven;
- ② It is elementary—we don't need to make statements about unobservable objects or unknowable distributions.
- ③ It is practical, and fully theory-consistent

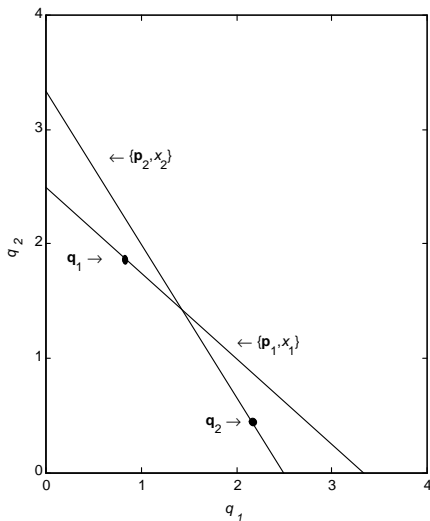
- 1 We partition the data into theory-consistent subsets.
- 2 We look for the partition that has the smallest number of groups.
Partitioning requires:
 - 1 Criteria for detecting differences between agents: Revealed Preference Restrictions
 - 2 A feasible computational approach: repeated subsampling

Revealed Preference Restrictions

Afriat's (1967) Theorem: subject to the caveat that preferences/technology are common across observations of choices, revealed preference (RP) restrictions:

- 1 are necessary and sufficient for the rationalizability of observed choices;
- 2 are inequality restrictions, so LP methods can be used to check whether or not they are satisfied;
- 3 can be used to check rationality, even with optimisation errors;
- 4 can be used in any optimisation context (firms, consumers, nations, etc)

Revealed Preference Restrictions



Revealed Preference (RP) restrictions (consumers)

Theorem. (Afriat (1967), Diewert (1973) and Varian (1982)); If the data satisfy the generalised axiom of revealed preference (RP) restrictions,

$$\mathbf{p}'_i \mathbf{q}_i \geq \mathbf{p}'_i \mathbf{q}_j, \mathbf{p}'_j \mathbf{q}_j \geq \mathbf{p}'_j \mathbf{q}_k, \dots \geq \mathbf{p}'_r \mathbf{q}_s \Rightarrow \mathbf{p}'_s \mathbf{q}_s \leq \mathbf{p}'_s \mathbf{q}_i \forall i, \dots, s \in I,$$

then there exists a concave, monotonic, continuous, non-satiated utility function $u(\mathbf{q})$ such that the data exactly solves the model

$$\max_{\mathbf{q}} u(\mathbf{q}) \text{ subject to } \mathbf{p}'_i \mathbf{q} = \mathbf{p}'_i \mathbf{q}_i \forall i \in I$$

- If $\{\mathbf{p}_t, \mathbf{q}_t\}_{t=1, \dots, T}$ relate to a *single individual over time* then this restriction is interpretable as a check for theoretical consistency and stability of a well-behaved set of preferences.
- If $\{\mathbf{p}_i, \mathbf{q}_i\}_{i=1, \dots, N}$ relate to a *cross section of individuals* then this restriction is interpretable as a check for theoretical consistency and commonality of a well-behaved set of preferences.

We turn Afriat's Theorem on its head: instead of assuming commonality and checking rationality, we assume rationality and assess heterogeneity.

- 1 Using the RP inequalities, we characterise the *minimum* amount of unobserved preference heterogeneity needed to rationalise the observed variation in consumer choice behaviour.
- 2 The *actual* number of types out there is not interesting—it is N .
- 3 The minimum number of types is both interesting and useful:
 - 1 useful for parsimonious modelling (eg, discrete-type macro-labour);
 - 2 useful for prediction (tightest predictions consistent with rationality);
 - 3 can be assessed in any context with optimising agents (firms, consumers, countries, etc)

Brute Force Approach

Say we observe a cross section (later, panel) of N agents $i = 1, \dots, N$, and we have observed choices X_i for each of them:

- 1 consider exclusive exhaustive groupings (from now on *partitions*) of the agents: do they satisfy rationality restrictions within groups?;
- 2 consider *all* partitions that satisfy within-group RP restrictions;
- 3 which has the smallest number of groups? what is that number? what differentiates the groups?

Brute Force Approach

- 1 Brute force approach answers our question exactly.
- 2 But, it is BIG job:
 - 1 for each partition:
 - 1 test each subset of the data for RP;
 - 2 write "pass" if all subsets pass RP test;
 - 2 look across passing partitions (comprised of mutually exclusive exhaustive RP-passing subsets of the data) to find the partition with the smallest number of groups;
- 3 this is 2^N partitions, each of which has as many as $N/2$ RP tests.

Non-Uniqueness of Partitions

Consider $N = 4$ — people are $\{1, \dots, 4\}$. Let 1 and 4 be *enemies*, who cannot be put in the same group.

Then, the minimum number of groups is 2:

$$\text{Admissible partitions} = \left\{ \begin{array}{l} [\{1, 2, 3\} \{4\}], \\ [\{1, 2\} \{3, 4\}], \\ [\{1, 3\} \{2, 4\}], \\ [\{1\} \{2, 3, 4\}], \end{array} \right\}$$

no unique minimizer partition; not equally heterogeneous.

This minimum number of types:

- 1 Brute force method is exact, but computationally burdensome—at least 2^N RP tests
- 2 Instead, we put bounds on the minimum number of types which
 - 1 can be bounded tightly on both sides;
 - 2 have associated partition(s), whose preferences can be characterised;
- 3 We can allow for optimisation error, panel data structures, other optimising contexts.

We have use 2 algorithms to find bounds.

- 1 They produce sharp, *two-sided bounds* on the number of groups and return a single partition, corresponding to the upper bound, satisfying within-group rationality.
- 2 It is based on repeated random ordering of the data.
- 3 There are $N!$ such orderings.
- 4 If you went through them all you would find the exact minimum number of types¹.
- 5 On each randomisation the bounds (weakly) crunch together.

Upper Bounds

- 1 loop over random orderings of the data, $b = 1, \dots, B$
 - 1 loop over the (randomly ordered) observations $i = 1, \dots, N$
 - 1 assign individuals to groups that don't fail RP
- 2 This uses at most BN^2 RP tests (a lot less than 2^N).
- 3 select the minimum number of groups over $b = 1, \dots, B$: this is an upper bound.

- 1 If 2 observations fail an RP test, then adding observations won't create a pass.
- 2 define a *group of enemies* as a group of observations such that no two can pass an RP test.
- 3 using random orderings of the data, find the biggest group of enemies possible—at least this many groups are needed.

- 1 We observe a cross section (panels are do-able, too) of N agents $I = \{1, \dots, N\}$.
- 2 We observe their budgets (including prices) and consumption choices $\{\mathbf{p}_i, \mathbf{q}_i\}_{i=1, \dots, N}$
- 3 Our Data: monthly milk purchases for 500 Danish households.
- 4 6 types of milk, with no unobserved quality variation (that's Denmark!).
- 5 Tons of heterogeneity in household demographics, some old, some young, some with kids, etc.
- 6 There is cross-sectional variation in prices that is not due to unobserved differences in product quality

TABLE 1: Descriptive Statistics

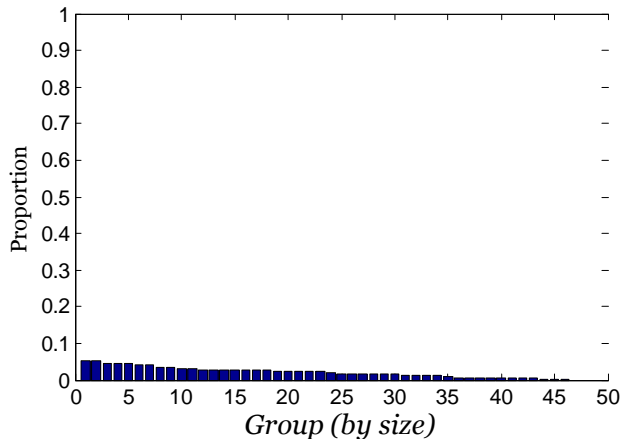
	Mean	Min	Max	Std. Dev
$\{\mathbf{w}_i\}_{i=1,\dots,500}$	Budget Shares			
Conventional Full Fat	0.1688	0	1.0000	0.3158
Conventional Semi-skimmed	0.4255	0	1.0000	0.4102
Conventional Skimmed	0.1521	0	1.0000	0.2932
Organic Full Fat	0.0374	0	1.0000	0.1394
Organic Semi-skimmed	0.0977	0	1.0000	0.2237
Organic Skimmed	0.1185	0	0.9951	0.2669
	Total Expenditure (DK)			
Total Expenditure	66.1986	4.8222	345.1279	58.5765
$\{\mathbf{p}_i\}_{i=1,\dots,500}$	Prices (DK litre)			
Conventional Full Fat	6.1507	3.3068	11.3289	0.4652
Conventional Semi-skimmed	5.4104	4.0919	7.9567	0.4304
Conventional Skimmed	5.1524	4.1619	6.2075	0.1814
Organic Full Fat	7.3335	6.1188	8.6597	0.1860
Organic Semi-skimmed	6.4968	5.0565	8.5374	0.2187
Organic Skimmed	6.2679	5.5312	7.9684	0.1501

$\{\mathbf{z}_i\}_{i=1,\dots,500}$	Demographics			
Singles $\{0, 1\}$	0.3260	0	1.0000	0.4692
Singles Parents $\{0, 1\}$	0.0420	0	1.0000	0.2008
Couples $\{0, 1\}$	0.3500	0	1.0000	0.4774
Couples with children $\{0, 1\}$	0.2300	0	1.0000	0.4213
Multi-adult $\{0, 1\}$	0.0520	0	1.0000	0.2222
Age (<i>Years</i>)	47.8600	18.0000	87.0000	15.5240
Male HoH $\{0, 1\}$	0.92	0	1.0000	0.27156

Partitioning on Observables

- 1 Demand analysis usually hinges preference heterogeneity on observables
- 2 So, we stratifying the data according to observables and run RP tests within groups.
- 3 We refine the stratification on observables until the RP test for commonality of preferences within types is satisfied.
 - 1 household structure (5 groups). No dice
 - 2 household structure x age of head (8 groups). No dice
 - 3 household structure x age of head x region (9 groups). No dice
 - 4 household structure x age of head x region x sex (2 groups). Yay!

Partitioning on Observables



Partitioning on Unobservables

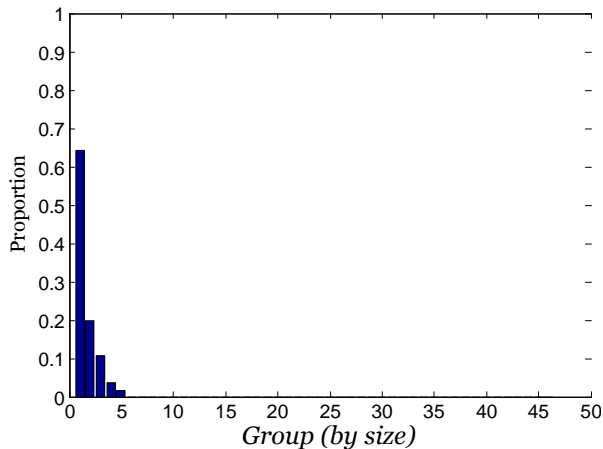
500 households—how many types?

Partitioning on Unobservables

500 households—how many types?

4 to 5

Partitioning on Unobservables: Upper Bound Partition



Who's in Which Group?

Given our "minimum number of groups" partition (with 5 groups), do observable covariates "explain" who is in which group?

- we totally expected a big "Yes"
- but, we got a big "No, not really".
- pseudo- R^2 in multinomial logit of group membership on observable covariates (excluding prices and budgets) is 4%.

Are the 5 Groups Very Different?

TABLE 2: Average Budget Shares Across Types

Group	Group <i>N</i>	Conventional Milk			Organic Milk		
		Full-fat	Semi	Skim	Full-fat	Semi	Skim
pooled	500	0.168	0.425	0.152	0.037	0.097	0.118
Type 1	321	0.160	0.496	0.143	0.024	0.075	0.100
Type 2	100	0.155	0.285	0.205	0.070	0.121	0.162
Type 3	53	0.239	0.351	0.074	0.044	0.144	0.147
Type 4	18	0.134	0.256	0.148	0.032	0.258	0.170
Type 5	8	0.292	0.195	0.357	0.128	0.017	0.009

TABLE 3: QAI Estimates of levels at median constraint

Group	Group N	Conventional Milk			Organic Milk		
		Full-fat	Semi	Skim	Full-fat	Semi	Skim
pooled	500	0.154 <i>0.018</i>	0.400 <i>0.024</i>	0.163 <i>0.017</i>	0.041 <i>0.008</i>	0.100 <i>0.013</i>	0.0142 <i>0.015</i>
group 1	321	0.155 <i>0.022</i>	0.434 <i>0.030</i>	0.173 <i>0.021</i>	0.020 <i>0.007</i>	0.085 <i>0.014</i>	0.133 <i>0.017</i>
group 2	100	0.153 <i>0.032</i>	0.287 <i>0.044</i>	0.194 <i>0.041</i>	0.089 <i>0.021</i>	0.092 <i>0.028</i>	0.184 <i>0.032</i>
group 3	53	0.195 <i>0.055</i>	0.330 <i>0.064</i>	0.091 <i>0.033</i>	0.070 <i>0.027</i>	0.130 <i>0.036</i>	0.184 <i>0.041</i>
group 4	18	0.084 <i>0.061</i>	0.171 <i>0.075</i>	0.295 <i>0.079</i>	0.052 <i>0.028</i>	0.209 <i>0.061</i>	0.190 <i>0.057</i>

Really Really?

TABLE 4: Nonparametric Estimates, averaged over all constraints

Group	Group N	Conventional Milk			Organic Milk		
		Full-fat	Semi-fat	Skimmed	Full-fat	Semi-fat	S
Average Levels							
group 1	321	0.157 <i>0.015</i>	0.465 <i>0.020</i>	0.158 <i>0.014</i>	0.027 <i>0.006</i>	0.078 <i>0.014</i>	
group 2	100	0.168 <i>0.022</i>	0.292 <i>0.035</i>	0.191 <i>0.029</i>	0.074 <i>0.013</i>	0.092 <i>0.015</i>	
group 3	53	0.201 <i>0.056</i>	0.362 <i>0.056</i>	0.073 <i>0.019</i>	0.034 <i>0.017</i>	0.128 <i>0.032</i>	
Average Semi-Elasticity wrt Expenditure							
group 1	321	-0.020 <i>0.022</i>	0.013 <i>0.028</i>	0.003 <i>0.016</i>	-0.001 <i>0.007</i>	-0.016 <i>0.009</i>	
group 2	100	-0.048 <i>0.028</i>	0.068 <i>0.061</i>	-0.084 <i>0.040</i>	-0.010 <i>0.013</i>	0.000 <i>0.025</i>	
group 3	53	0.028 <i>0.082</i>	-0.074 <i>0.094</i>	-0.029 <i>0.027</i>	-0.002 <i>0.017</i>	-0.054 <i>0.035</i>	

Heterogeneity in Preferences for Milk

- 1 We can *completely* explain the variation of observed behaviour with variation in budget constraints and 4 or 5 preference maps (ie. ordinal utility functions).
- 2 The groupings are not very related to observed characteristics—the primary heterogeneity here is *unobserved*.
- 3 The groups found by our upper bound algorithm are very different from each other.
 - 1 They differ by more than just level effects, so unobserved preference heterogeneity may not act like ‘error terms’ (or fixed effects) in regression equations (panel models).

- Same 500 households as before, but use a sequence of up to 24 months of milk consumption data for each household.
- "break people into pieces" each of which satisfies RP tests.
- The number of groups needed to completely rationalise these data **is at least 12 and not more than 31**.
- fixed effects in this context would imply 500 groups.
 - 1 This is at least 15 times as many groups as are really necessary, and therefore is radically overspecified.
 - 2 Blundell, Duncan and Pendakur (1998) show that fixed effects in budget shares tough to rationalise
 - 3 Above, we showed that *both* levels *and* derivatives of budget-share equations vary across groups.
 - 4 So, fixed effects are both too much and too little.

- 1 One can introduce a parameter $e \in [0, 1]$ (the Afriat efficiency parameter), and modify the Afriat inequalities such that

$$e\mathbf{p}'_t\mathbf{q}_t \geq \mathbf{p}'_t\mathbf{q}_s \Leftrightarrow \mathbf{q}_t R_e^0 \mathbf{q}_s.$$

- 2 The weaker RP restriction is then

$$\mathbf{q}_j R_e \mathbf{q}_i \Rightarrow e\mathbf{p}'_i\mathbf{q}_i \leq \mathbf{p}'_i\mathbf{q}_j$$

where R_e is the transitive closure of R_e^0 .

- 3 The interpretation of e is as the proportion of the consumer's budget which they are allowed to waste through optimisation errors.
- 4 Given a value of e , one can find the minimum number of groups as we have done. Set e low enough, and that minimum number is always 1.

Optimisation Error

e	Number of Types	
	Lower bound	Upper bound
0.78	1	1
0.80	1	2
0.85	1	2
0.90	2	3
0.95	3	4
1.00	4	5

- For $e < 0.781$, only 1 type is needed.
- For $e > 0.900$, more than 1 type is needed.

Revealed Preference in other Contexts

Utility maximisation: Afriat (1967), Diewert (1973) and Varian (1982); Profit maximisation and cost minimisation by perfectly competitive and monopolistic firms: Hanoch and Rothschild (1972); The strong rational expectations hypothesis: Browning (1989); Expected utility theory: Bar-Shiva (1992); Collective Household models: Cherchye, De Rock and Vermuelen (2007, 2010); Firm investment behaviour: Varian (1983b); Characteristics models: Blow et al (2008); useful functional restrictions: additive/weak/latent separability, homotheticity, returns to scale etc. These RP restrictions generally

- 1 only involve inequality restrictions on observables;
- 2 exhaust *all* of the nonparametric empirical implications of the theory;
- 3 can be used to assess heterogeneity in all these contexts

Firms (Farms)

- We can do the same with firm data.
- 281 Danish Farms observed in 1990. Detailed annual accounts of variable costs and earnings for each production line with corresponding accounts measures of most inputs and outputs.
- We measure five outputs {milk, two types of beef, and two types of crops} and we observed 46 inputs - like fodder, cattle, fertiliser, pesticides, and the services from labour, land, building and machine capital.
- Here, we are interested in unobserved *technological* heterogeneity.
- The minimum number of technologies **is at least 3 and not more than 4.**

- 1 Since Becker and Stigler's assessment, we have learned much about preference heterogeneity that is correlated with observables.
- 2 But, it seems that the more important kind of heterogeneity is driven by unobservables.
- 3 Our results suggest that models which use a small number of heterogeneous types may in fact be dealing with unobserved heterogeneity in a sufficient fashion.
 - 1 E.g., macro-labour models, education choice models, and a vast number of empirical marketing models may be okay.
 - 2 Models like Lewbel and Pendakur (2009), in which unobserved preference heterogeneity is captured by a multidimensional continuum of unobserved parameters could be overkill.