Who are the Indian Middle Class? A Mixture Model of Class Membership Based on Durables Ownership

Sudeshna Maitra
York University & The Conference Board

2008 World Congress on National Accounts and Economic Performance Measures for Nations, Washington D.C.
May 12 – 17, 2008

Why study the middle class?

- Countries with a larger middle class tend to have higher growth rates (Easterly (2001))
- The middle class is the "backbone of the market economy and democracy in most advanced societies" (Birdsall, Graham and Pettinato (2000))
- 'Middle class values' emphasize human capital accumulation and savings (Banerjee and Duflo (2007))
- New entrepreneurs emerge from the middle class
- The middle class consumer demands quality consumer goods and is willing to pay a higher price for the same – provides 'big push' to investment (Murphy, Schleifer, Vishny (1989)

'The Great Indian Middle Class'

- India's impressive growth performance since the 1990s
- Focus on improved living standards of Indians
- The fortunes of the emerging middle class have received much attention in the media
- The middle class are the key to sustainable growth in India
  - They are ideally placed to partake of the 'trickle-down' benefits of growth and respond to economic incentives and policy (Sridharan (2004))
- India's population is one-sixth of the world's; hence the Indian middle class constitutes a sizeable portion of the global workforce and consumer body
- The middle class has been traditionally perceived as
  - Lying between the very poor and the very rich: very wide definition (who are the 'very poor' and the 'very rich'?)
  - an educated section of society seeking white-collar jobs

Existing Approaches to Identifying the Middle Class

- Very few attempts in the economic literature to rigorously identify and measure the size of the middle class

- Existing definitions of the middle class:
  - Easterly (2001): middle class lies between 20th and 80th percentile of the consumption distribution
  - Birdsall et al (2000): middle class lies between 75% and 125% of median income
  - Banerjee and Duflo (2007), Sridharan (2004), NCAER (2005), IBEF (2005), McKinsey (2007): Use income (or expenditure) cutoffs to define different classes

- Results are sensitive to the researcher's notion of who the middle class are

- Some of these definitions are 'size-invariant'

What's new here?

- Propose a robust method for identifying the middle class: the Mixture Model
- Define classes based on their consumption standards, specifically ownership of 12 durable goods
- The classes are identified on the basis that their durable ownership patterns are *different*
- Advantages of the approach:
  - No arbitrary assumptions on income or expenditure requirements (viz. percentiles, cutoffs) for determining who constitute the different classes
  - Comparability across countries and societies
- Estimates are obtained for:
  - The proportions of households in the lower, middle and upper class
  - The durable ownership density of each class
  - Class-membership probabilities of households owning a particular number of durable goods

Road Map

- Methodology
  - Data
  - The Mixture Model & EM Algorithm
- Results
- Robustness checks

Methodology: Data

- National Sample Survey (NSS) Data, 55th Round, 1999-2000
- Urban Sub-sample: 48,924 households
- 12 durable goods
  - 5 Recreational goods: record player, tape/CD player, radio, TV, VCR/VCP
  - 4 Household goods: fan, AC, washer, refrigerator
  - 3 Transport goods: bike, motor bike/ scooter, car

Methodology: Why these 12 goods?

- Goods must have the following characteristics:
  - Non-necessary items
  - Indicate affluence

- Senauer & Goetz (2003): Major Household Durable Goods (Peru, National Living Standards Measurement Survey, 2000)
  - Common to S & G:
    - Radio, Refrigerator, Washing Machine, Automobile, TV
  - In addition, S & G included:
    - Black & White TV, Color TV, Wired Telephone, Cellular Telephone, Computer, Videocassettes
  - S & G omitted:
    - VCR/VCP, Record player, Tape/CD player, AC, Fan, Bicycle, Motor Bike (automobile?)

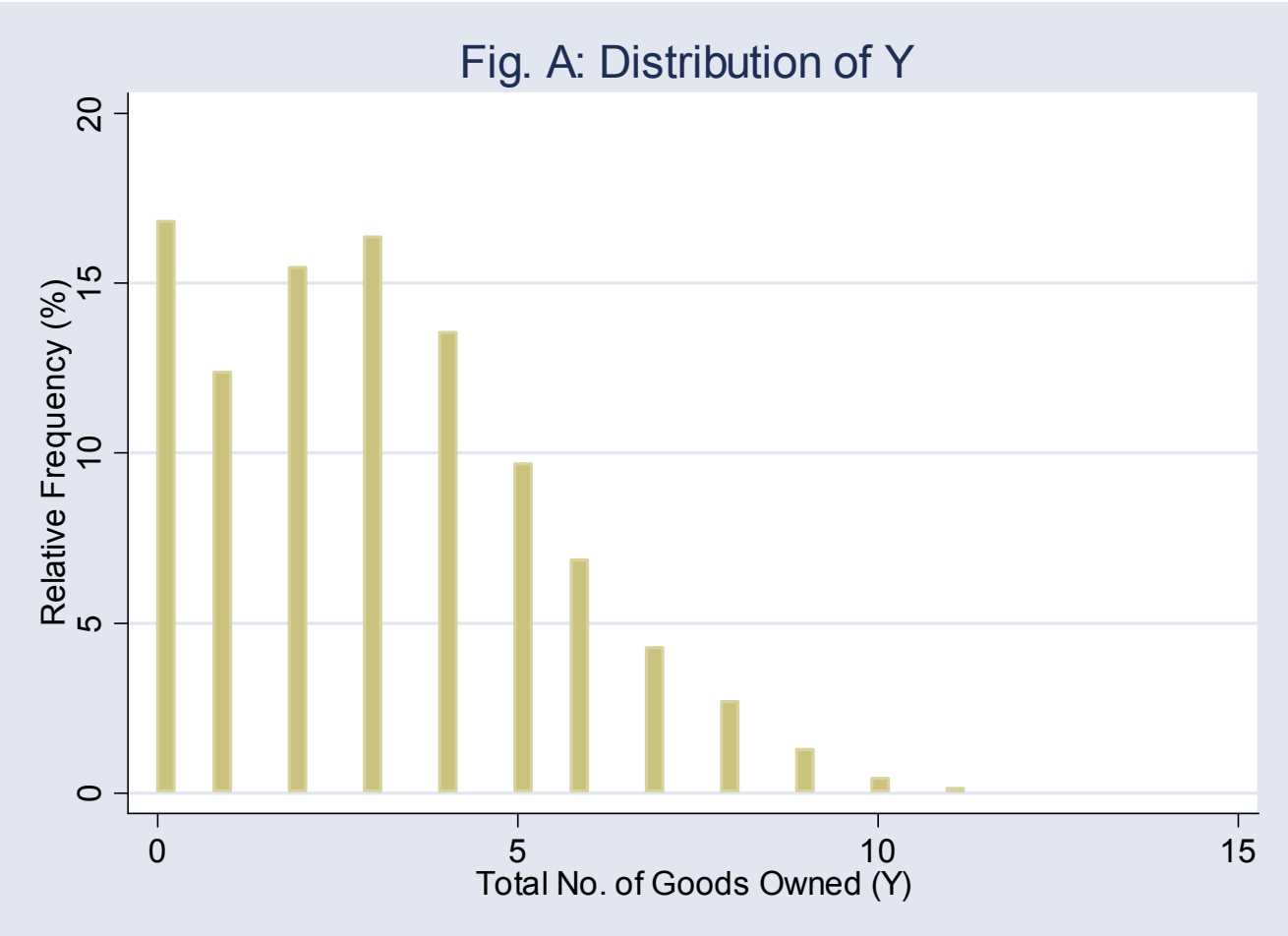
Methodology: Variable Definition

- Define 'Ownership' of a durable Z
  - = 1 if at least one piece of the durable Z is in use at the time of survey
  - = 0, otherwise

- Let *Y* denote the **total number of durables** 'owned' by a household at the time of survey (out of 12 goods)

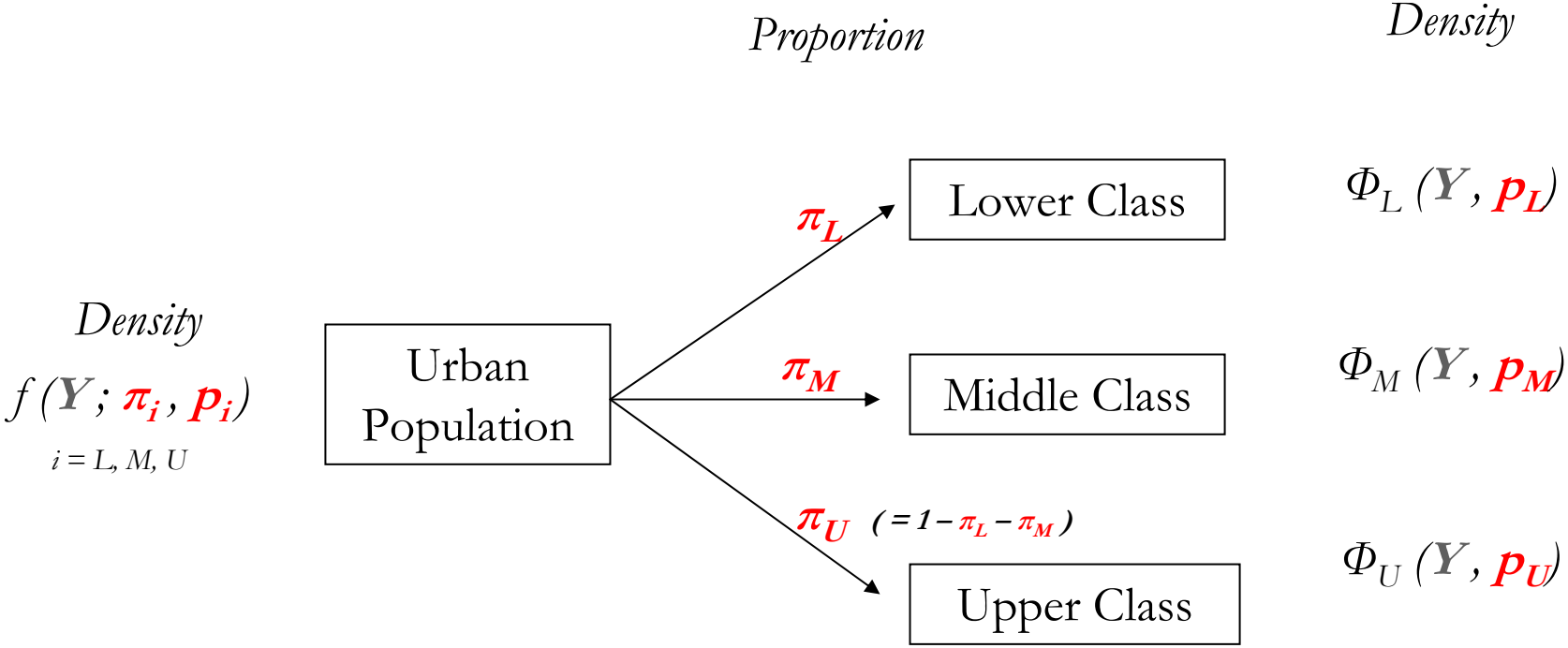
- *Y* is the sum of 12 binary variables, each denoting the ownership of a durable good ($0 \leq Y \leq 12$)

Durable Ownership (*Y*) in NSS, 1999-00 Urban Sub-sample



Methodology: The 3-Component Mixture Model

The density of total durable ownership (Y) in the population is a 'mixture' of the densities of Y for each class



$$f(y; \pi_L, \pi_M, p_L, p_M, p_U) = \pi_L \Phi_L(y; p_L) + \pi_M \Phi_M(y; p_M) + \pi_U \Phi_U(y; p_U)$$

Observed in data ; **Parameter to be estimated**

Methodology: Assumptions

- There are three classes in the population: lower (*L*), middle (*M*) and upper (*U*)
- Each good is owned independently with a fixed probability p_i, where *i* represents the class to which the household belongs
- Hence *Y* (the total of 12 goods owned by a class-*i* household) follows a binomial distribution with parameters $(12, p_i)$
- $p_L < p_M < p_U$
- E.g. Probability that a class-*i* household owns '*x*' of the 12 durable goods

$$= Pr(Y = x \mid \text{household belongs to Class } i)$$

$$= {}^{12}C_x (p_i)^x (1 - p_i)^{12-x}$$

Methodology: The Mixture Model

- The probability of observing a value $Y = y$ in a random sample is

$$f(y; \pi_L, \pi_M, p_L, p_M, p_U) = \pi_L \Phi_L(y; p_L) + \pi_M \Phi_M(y; p_M) + \pi_U \Phi_U(y; p_U)$$

- Hence the likelihood of picking ***an i.i.d sample*** of n households with ownerships $(y_1, y_2, \ldots, y_n)$ is

$$F(y_1, y_2, \ldots, y_n; \pi_L, \pi_M, p_L, p_M, p_U) =$$
$$\Pi_j [\pi_L \Phi_L(y_j; p_L) + \pi_M \Phi_M(y_j; p_M) + \pi_U \Phi_U(y_j; p_U)]$$

(j subscripts households, $j = 1, 2, \ldots, n$)

Observed in data ; **Parameter to be estimated**

Methodology: The Mixture Model

- Log Likelihood

$$Log\ F(y_1, y_2, \dots, y_n; \pi_L, \pi_M, p_L, p_M, p_U)$$

$$= \sum_j log\,[\,\pi_L\,\Phi_L(y_j; p_L) + \pi_M\,\Phi_M(y_j; p_M) + \pi_U\,\Phi_U(y_j; p_U)\,]$$

- Maximum Likelihood Estimation (M.L.E.): Find parameters $\{\pi_L, \pi_M, p_L, p_M, p_U\}$ that maximize the above log likelihood

- Closed-form maximands are hard to obtain due to the log of a sum in the log likelihood

Observed in data ; **Parameter to be estimated**

Methodology: The EM Algorithm

- Suppose that class memberships of households, $\{\Delta_L, \Delta_M\}$, are known
  - $\Delta_{Lj} = 1$ if the household j belongs to class L, 0 otherwise
  - $\Delta_{Mj} = 1$ if the household j belongs to class M, 0 otherwise

  $(\Delta_{Uj} = 1 - \Delta_{Mj} - \Delta_{Uj})$

- Then the likelihood of observing $\{y_1, \ldots, y_n\}$ can be written as:

$$F_{EM}(y, \Delta_{Lj}, \Delta_{Mj}; \pi_L, \pi_M, p_L, p_M, p_U) =$$

$$\Pi_j \{\pi_L \Phi_L(y; p_L)\}^{\Delta_{Lj}} \{\pi_M \Phi_M(y; p_M)\}^{\Delta_{Mj}} \{\pi_U \Phi_3(y; p_U)^{\Delta_{Uj}}\}$$

Observed/ Assumed to be observed ; **Parameter to be estimated**

Methodology: The EM Algorithm

- The log likelihood is then

$$\text{Log } F_{EM}(y, \Delta_{Lj}, \Delta_{Mj}; \pi_L, \pi_M, p_L, p_M, p_U) =$$

$$\sum_j [\Delta_{Lj} \log\{\pi_L \Phi_L(y_j; p_L)\} + \Delta_{Mj} \log\{\pi_M \Phi_M(y_j; p_M)\} + \Delta_{Uj} \log\{\pi_U \Phi_3(y_j; p_U)\}]$$

- Sum of logs makes it easy to find closed-form expressions for parameters $\{\pi_L, \pi_M, p_L, p_M, p_U\}$ that maximize the EM log likelihood
- But $\{\Delta_1, \Delta_2\}$ are unknown!
- The EM algorithm computes expectations of $\{\Delta_1, \Delta_2\}$ from the data, plugs them into the EM log-likelihood and maximizes it

Observed/ Assumed to be observed ; **Parameter to be estimated**

Methodology: The EM Algorithm

1. Start with initial guesses for parameters:
   $\{\pi^{(0)}_1, \pi^{(0)}_2, p^{(0)}_1, p^{(0)}_2, p^{(0)}_3\}$

2. **The Expectation (E) Step**: at the k^{th} step, compute the expected value of class-i membership of household j, conditional on the data
   $\gamma^{(k)}_{ij} = E\left(\Delta_{ij} / y_1, \ldots, y_n; \pi^{(k-1)}_1, \pi^{(k-1)}_2, p^{(k-1)}_1, p^{(k-1)}_2, p^{(k-1)}_3\right)$

3. **The Maximization (M) Step**: at the k^{th} step, compute the parameters by maximizing the EM log likelihood function (substitute $\gamma^{(k)}_i$ for $\Delta^{(k)}_i$ in the EM log-likelihood function)

4. Repeat the **E** and **M** steps until convergence is obtained

Interpretation of Estimates

- Unknown parameters*: { $\pi_L, \pi_M, p_L, p_M, p_U$ }
- π_i : Proportion of class *i* in the population
- p_i : The probability that a household belonging to class *i* will own a representative good
- By definition of the classes, $p_L < p_M < p_U$
- γ_{ij} : (Cond.) Probability that household *j* belongs to class *i*

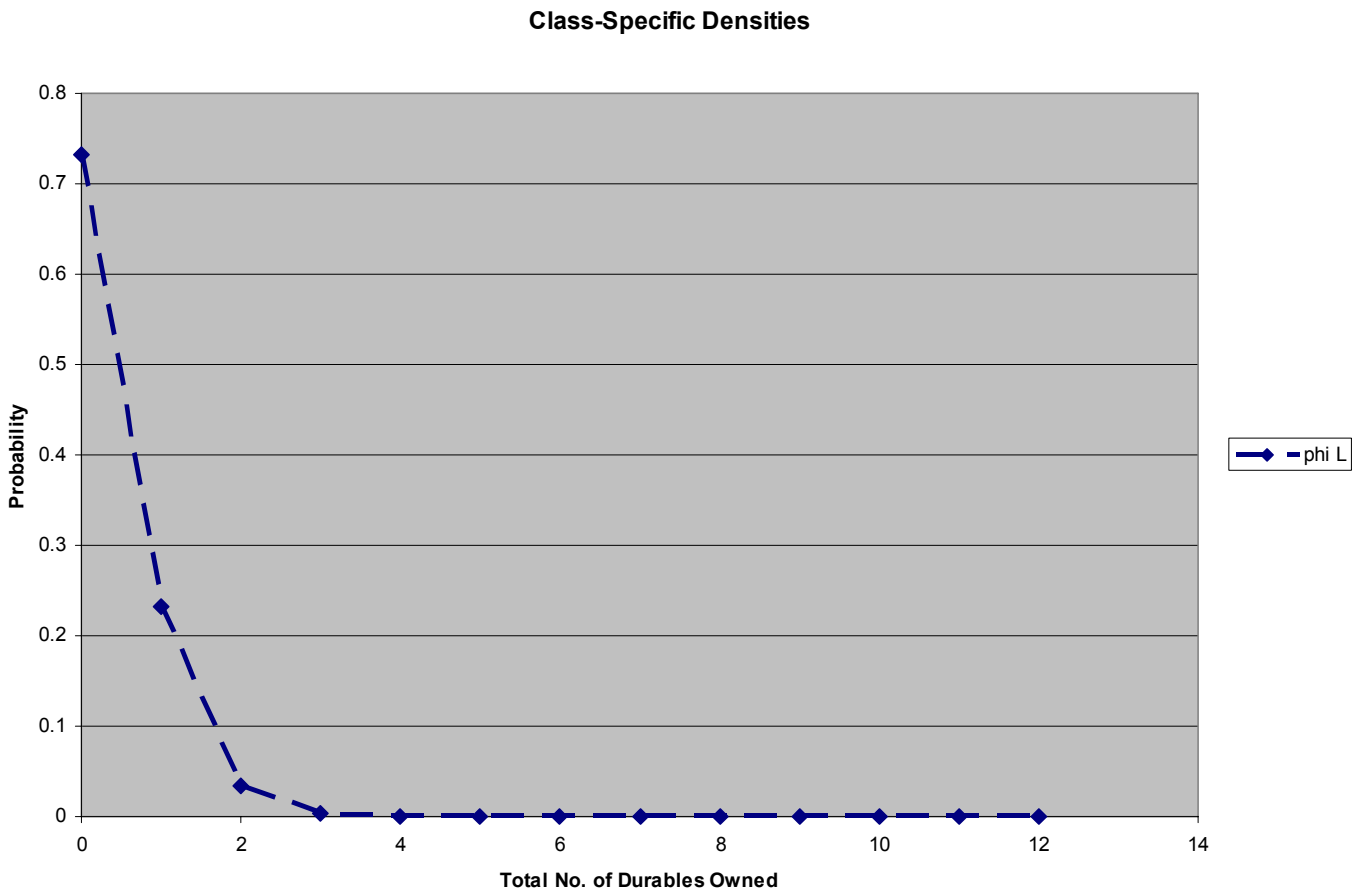
* $\pi_U = 1 - \pi_L - \pi_M$

Results: EM Estimates of Mixture Model Parameters

Class	Share of Urban Population (π_i)	Probability of Owning a Durable (p_i)	Mean No. of Durables Owned ($12p_i$)
Lower	0.2034 (0.005)	0.0257 (0.002)	0.3084
Middle	0.6162 (0.005)	0.251 (0.003)	3.012
Upper	0.1804 (0.006)	0.5249 (0.004)	6.2988

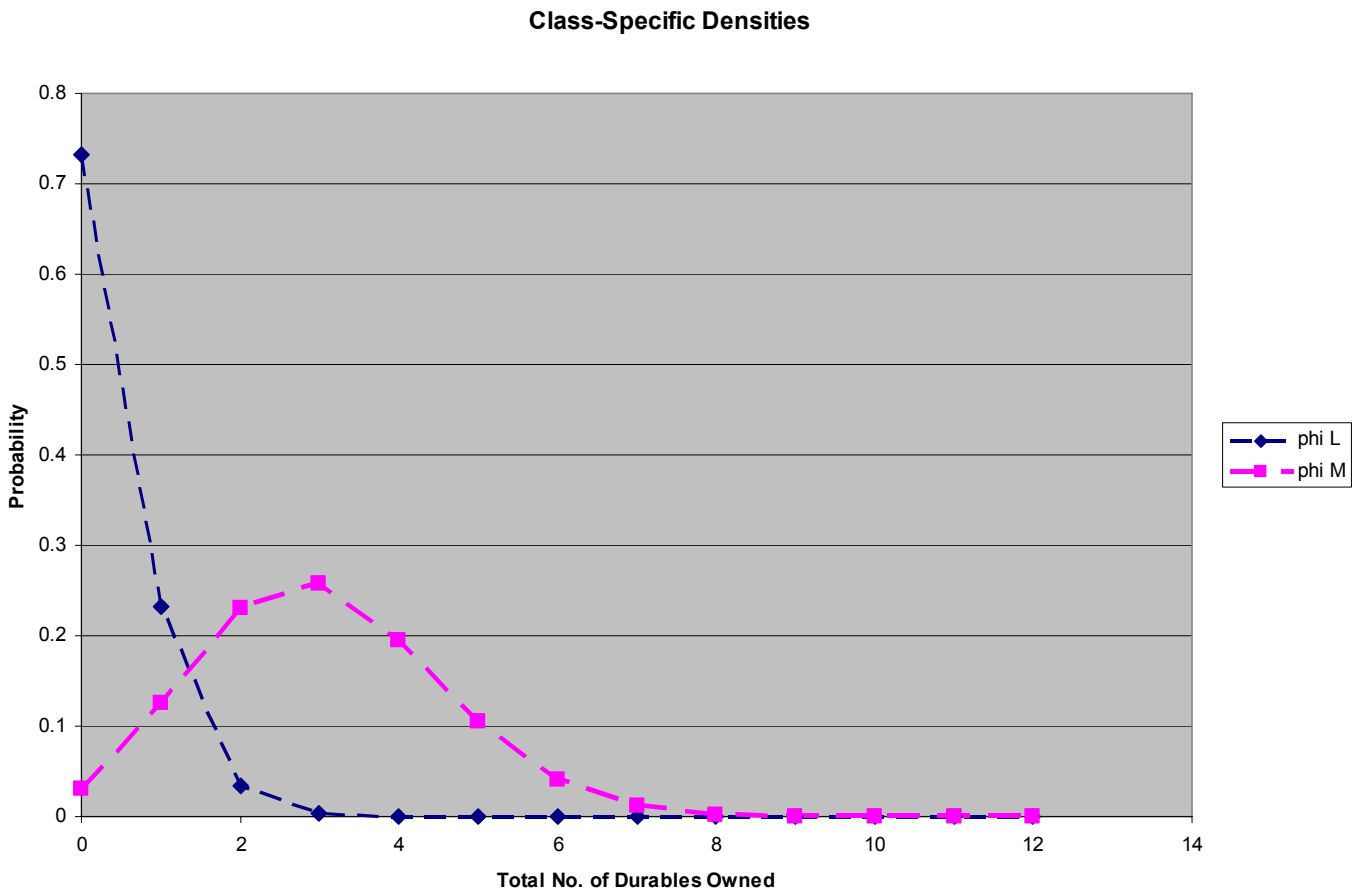
Standard Errors in Parentheses

Results: Estimated Densities $\Phi_i(Y; p_i)$



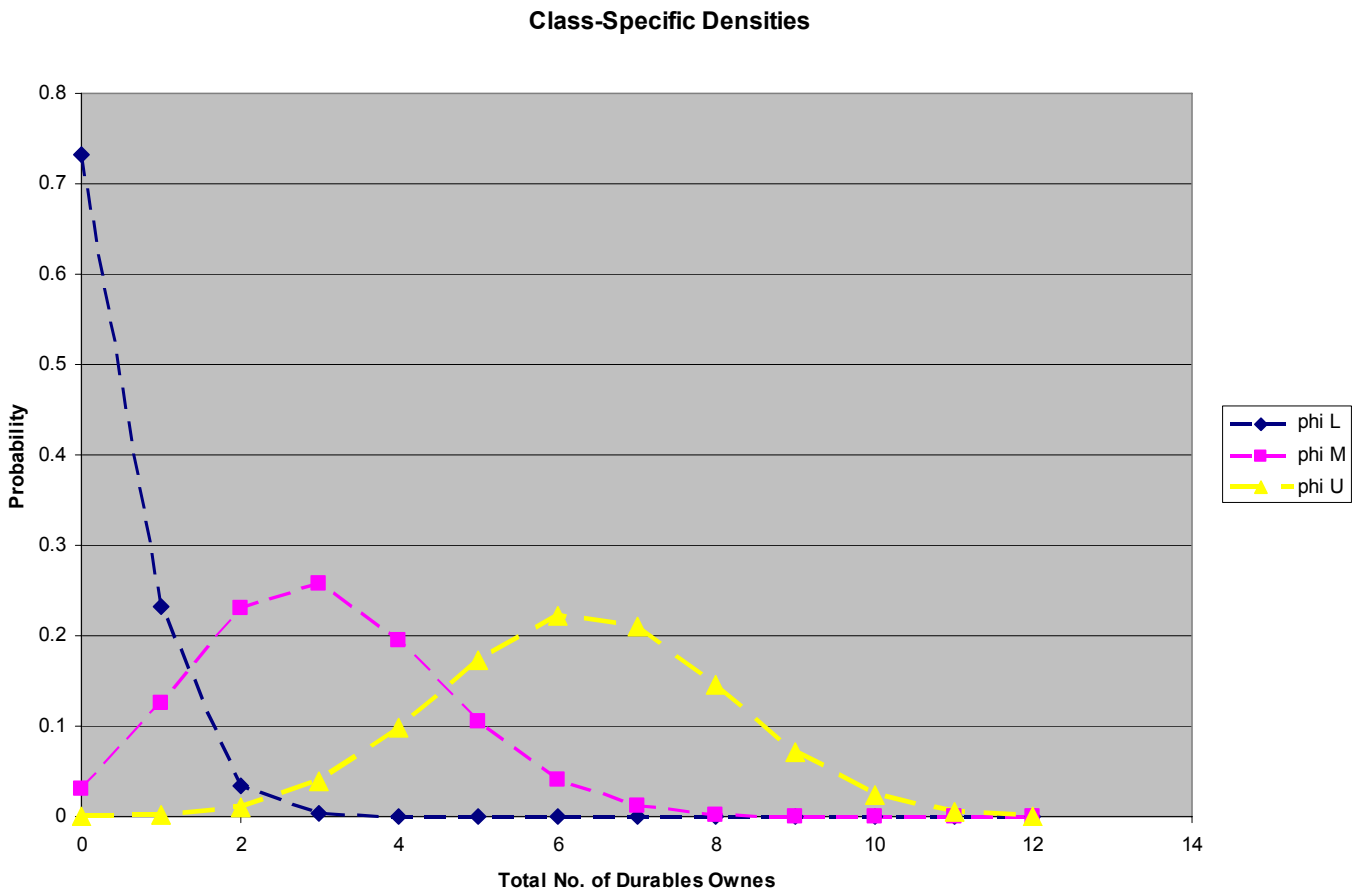
Estimates: $p_L = 0.0257$

Results: Estimated Densities $\Phi_i(Y; p_i)$



Estimates: $p_L = 0.0257$, $p_M = 0.251$

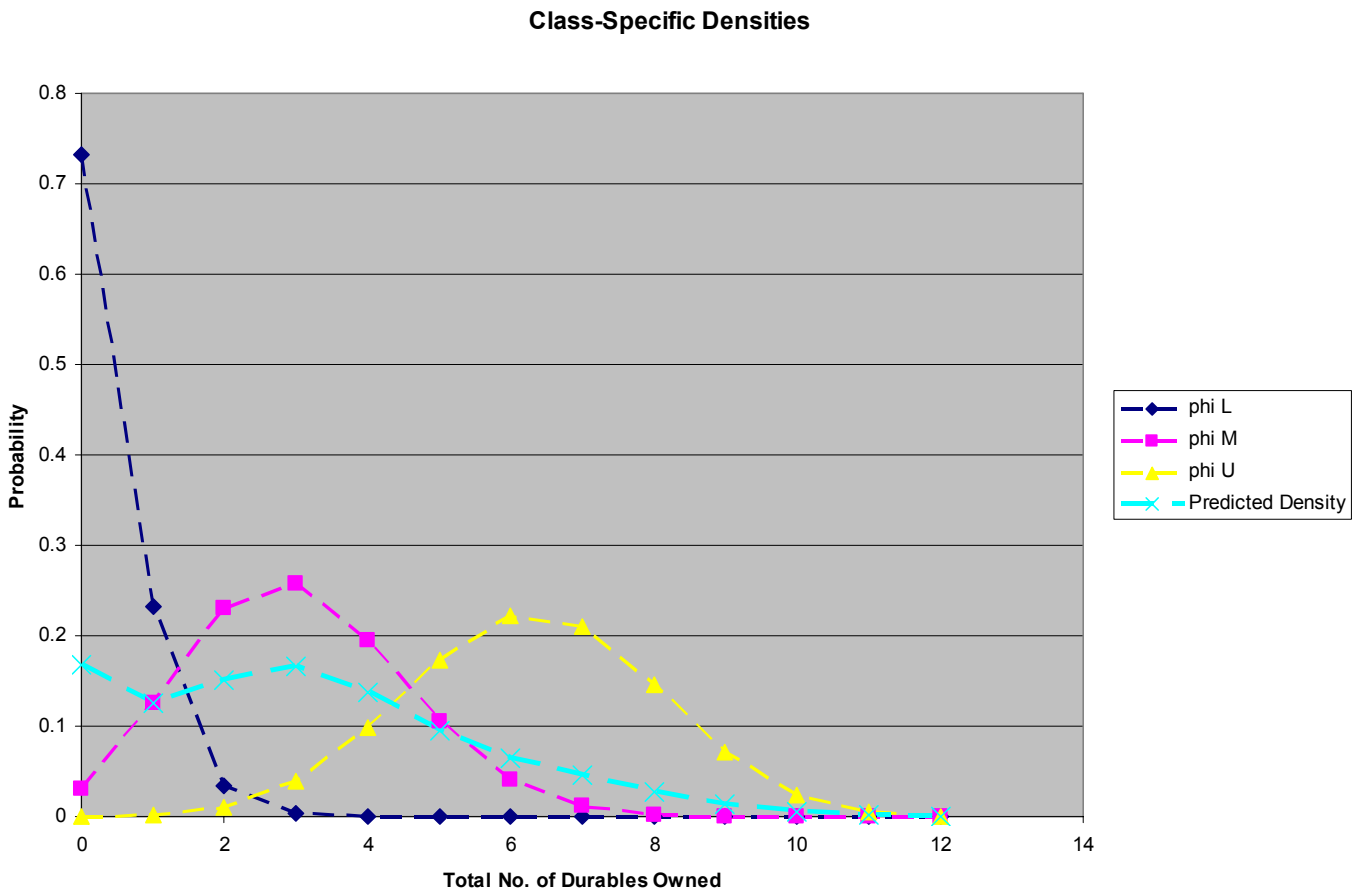
Results: Estimated Densities $\Phi_i(Y; p_j)$



Estimates: $p_L = 0.0257$, $p_M = 0.251$, $p_U = 0.5249$

Results: Predicted Density

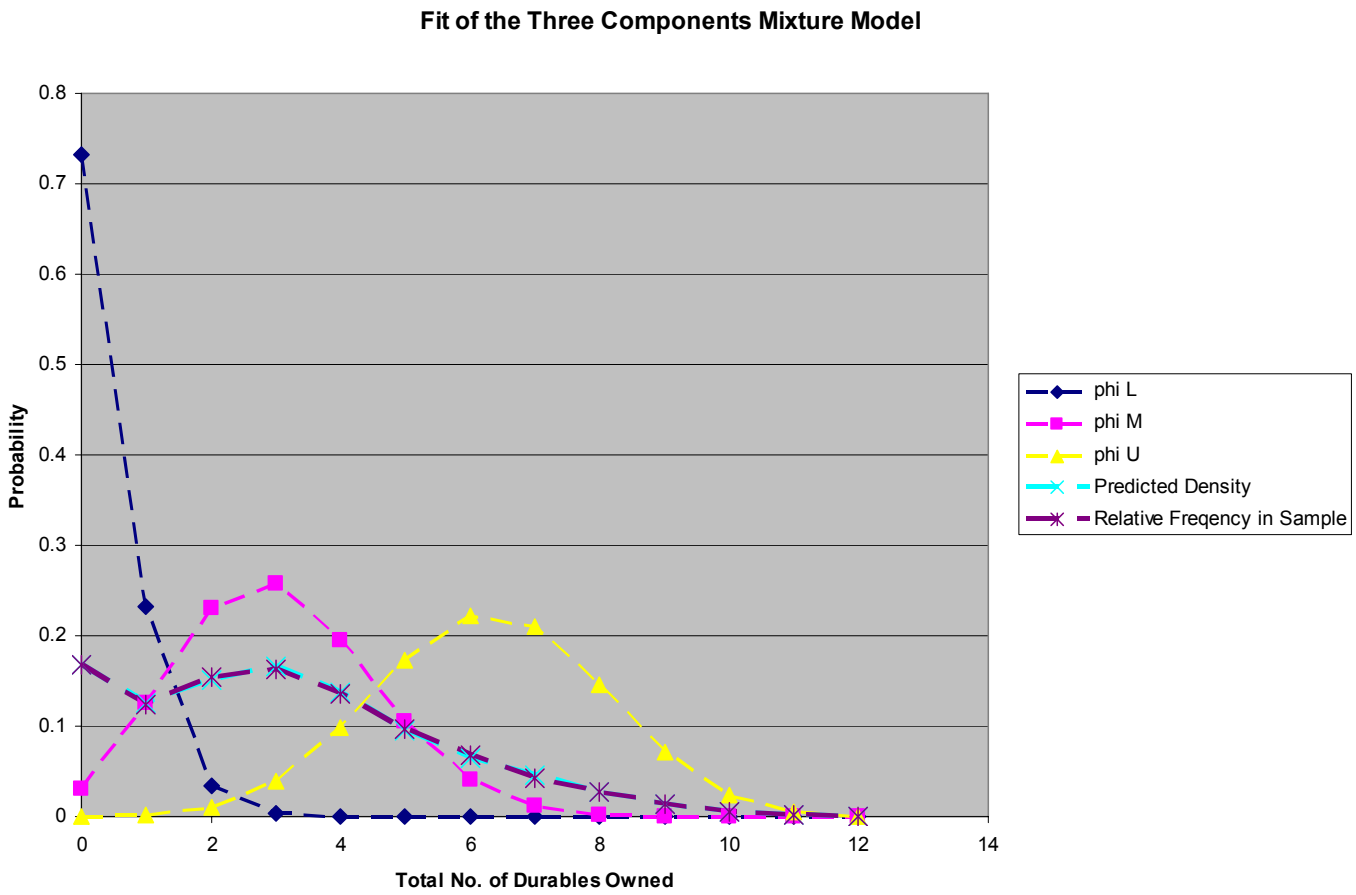
$$\textbf{Predicted Density} = \pi_L \Phi_L(y; p_L) + \pi_M \Phi_M(y; p_M) + \pi_U \Phi_U(y; p_U)$$



Estimates: $p_L = 0.0257$, $p_M = 0.251$, $p_U = 0.5249$, $\pi_L = 0.2034$, $\pi_M = 0.6162$, $\pi_U = 0.1804$

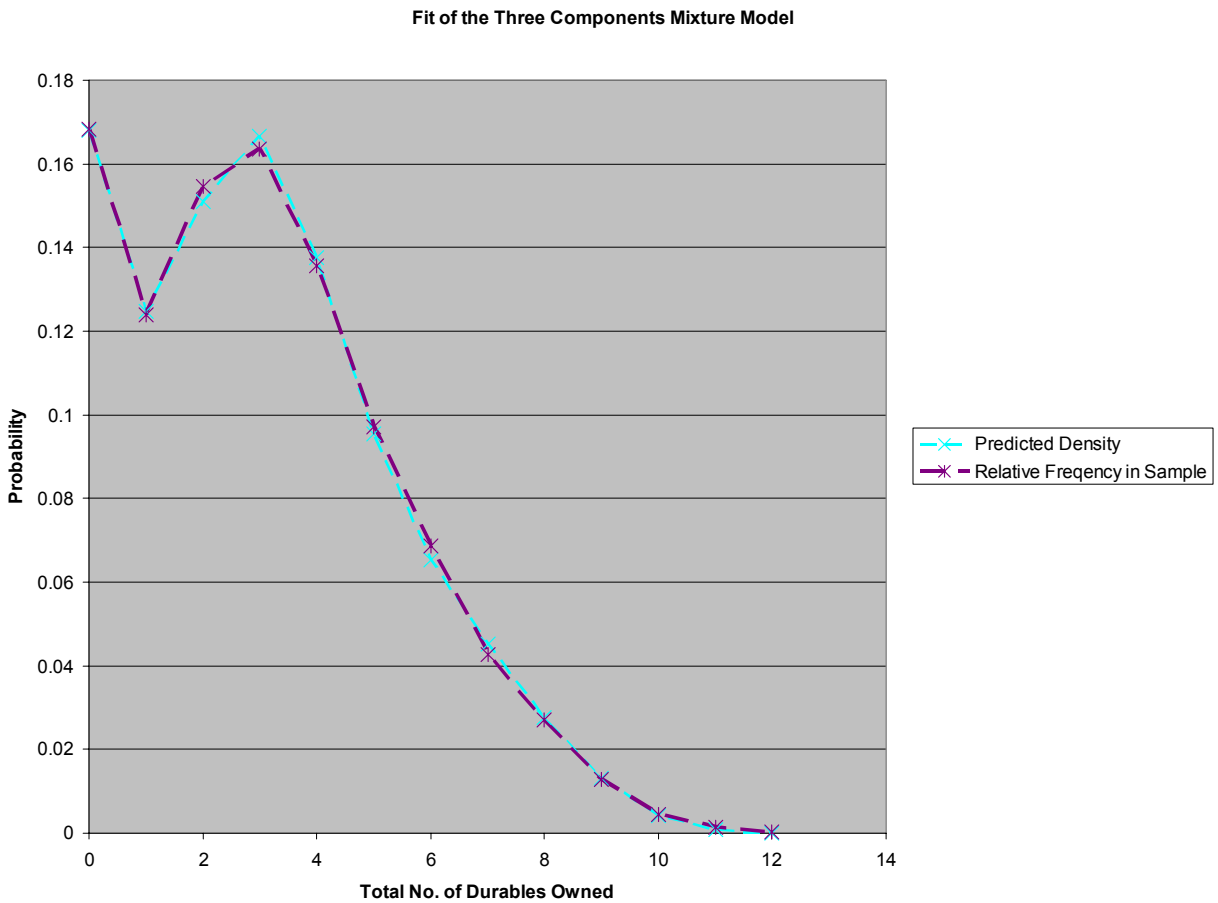
Results: Fit of the 3-Components Model

$$\textit{Predicted Density} = \pi_L \Phi_L(y; p_L) + \pi_M \Phi_M(y; p_M) + \pi_U \Phi_U(y; p_U)$$

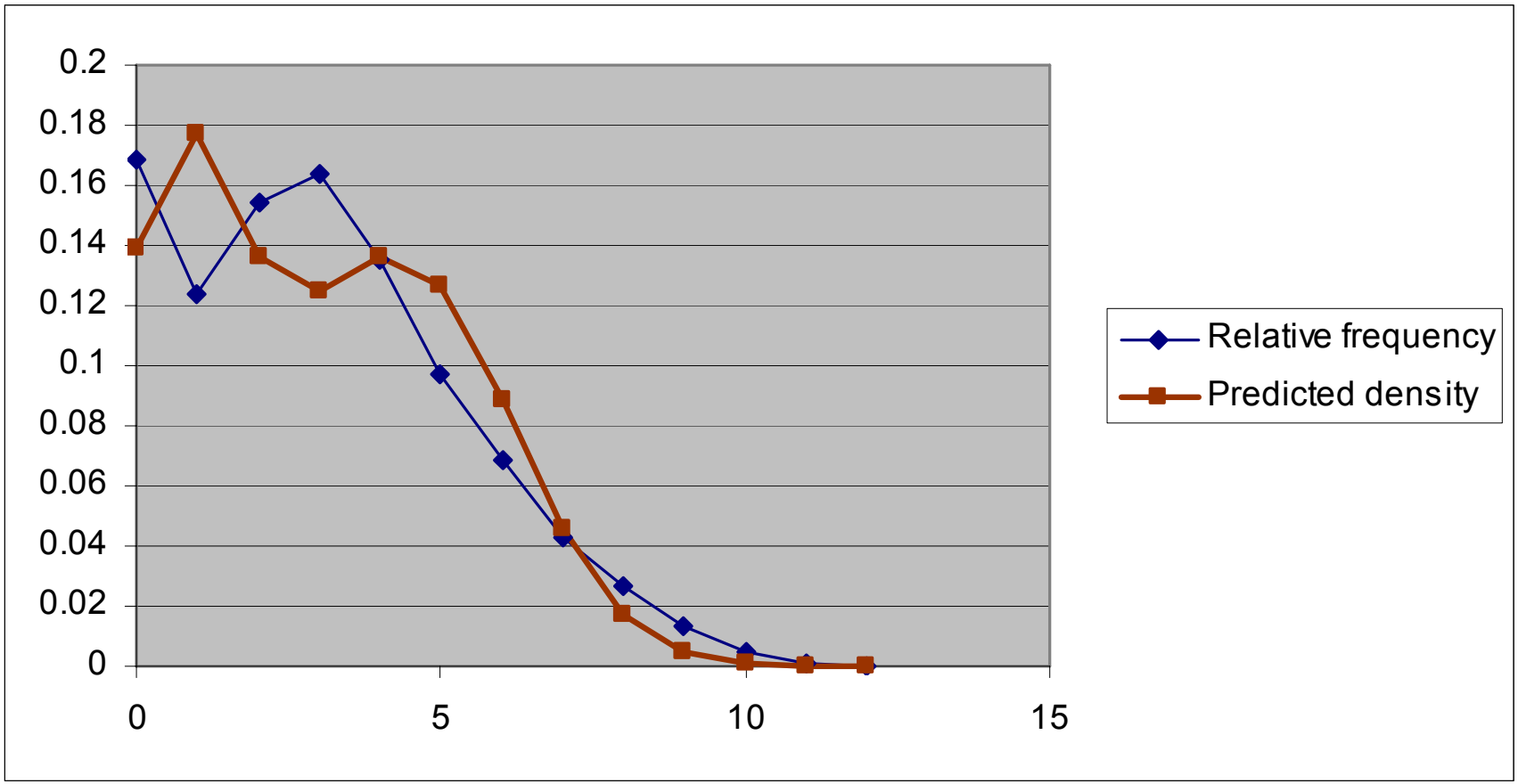


Results: Fit of the 3-Components Model

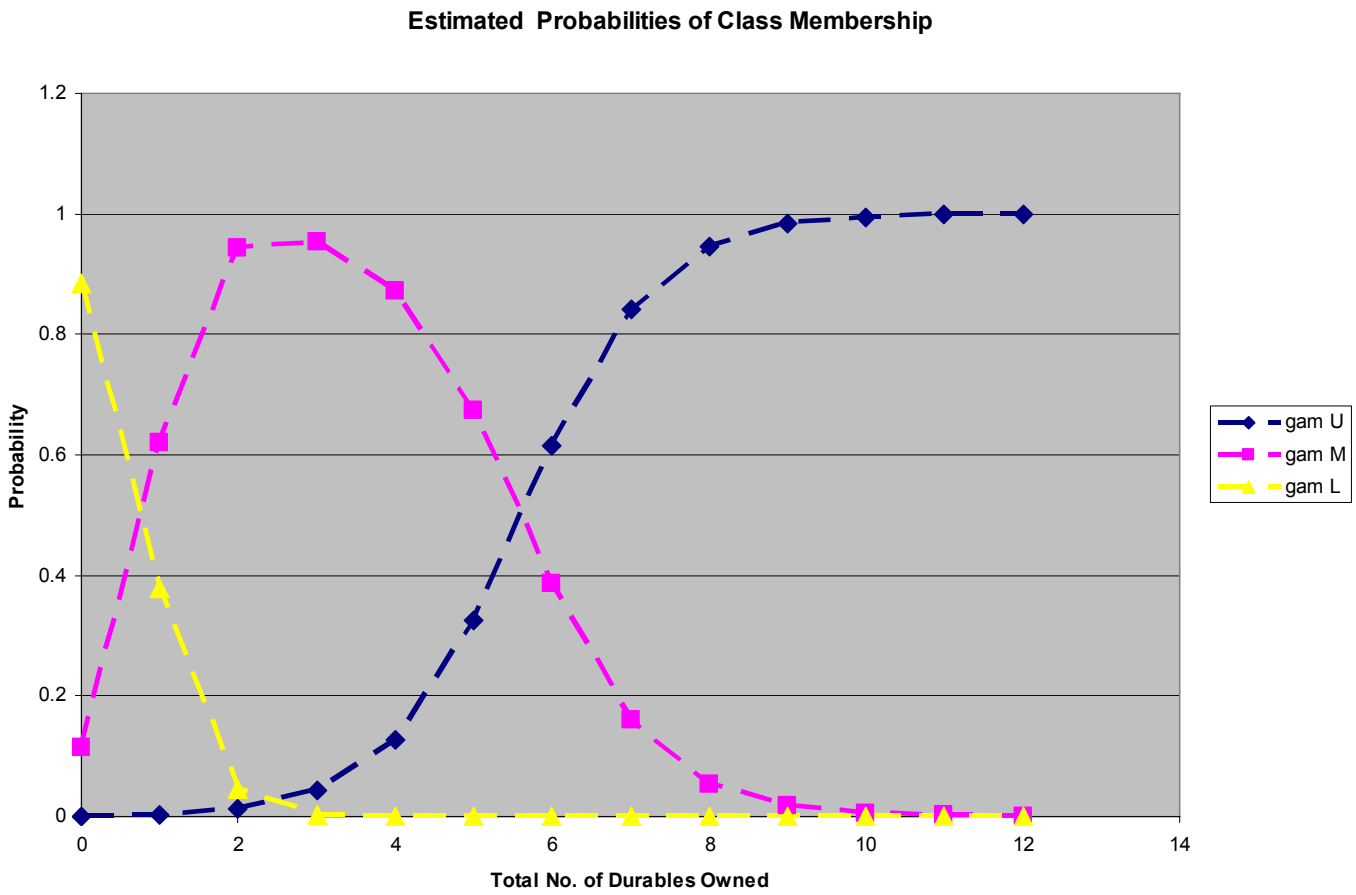
$$\textbf{Predicted Density} = \pi_L \Phi_L(y; p_L) + \pi_M \Phi_M(y; p_M) + \pi_U \Phi_U(y; p_U)$$



Compare: Fit of a 2-Component Mixture Model



Results: Probabilities of Class Membership, γ_i



Class Characteristics

- Randomly assign households to the 3 classes so as to be consistent with estimated densities and proportions

- Look at household characteristics, by assigned class
  - Per capita monthly expenditure

Socioeconomic Characteristics, by Assigned Class

Class	Per Capita Monthly Household Expenditure									Other Household Characteristics		
	Mean	Std. Dev.	Min.	Max.	Percentiles					Avg. No. of Meals Per Day Per Person	Proportion of Literate Household Members	Mean Household Size
					25	50	75	90	99			
Lower	791.2604 [$54.35]	859.1093 [$59.01]	17 [$1.17]	50528 [$3470.81]	423 [$29.06]	625 [$42.93]	981 [$67.39]	1421 [$97.61]	2791.43 [$191.75]	2.34	0.64	3.97
Middle	961.7854 [$66.07]	1772.39 [$121.75]	49 [$3.37]	205987 [$14,149.40]	532 [$36.54]	762 [$52.34]	1140 [$78.31]	1663 [$114.23]	3485 [$239.39]	2.38	0.77	4.65
Upper	1469.571 [$100.95]	1109.971 [$76.24]	224 [$15.39]	35612 [$2446.22]	842 [$57.84]	1229 [$84.42]	1777 [$122.06]	2490.6 [$171.08]	5390.08 [$370.25]	2.41	0.88	5.12

Per capita Monthly Household Expenditure: Percentiles

Percentile	10	20	30	40	50	60	70	80	90	99
Value	392	490	584	686	801	940	1120	1377	1815	3799.56

PPP US$ values in square brackets; converted at INR 14.558 per dollar (WHO, PPP, 2000)

Current vs. Previous Estimates

- Current estimates suggest that the Indian urban middle class
  - Has lower per capita expenditure than previously assumed
  - Is larger in size than previously estimated

Current vs. Previous Estimates

- NCAER study defines:
  - 'Middle Class': annual household income between Rs. 200,000 – 1000,000
  - 'Aspirers': annual household income between Rs. 90,000 – 200,000
- Here, median annual household incomes (at an average savings rate of 28% (McKinsey (2007)) are obtained to be:
  - Upper class: Rs. 104,465
  - Middle Class: Rs. 58,420
  - Lower Class: Rs. 41,354
- Banerjee and Duflo's cutoffs (daily PCE $2-$4 and $6-$10) are closer to current estimates though still on the high side (median daily PCE of the middle class = $1.9 and the 99th percentile is $8.50)
- Using Easterly's (2001) definition would exclude 20% of upper middle class members in the current analysis
- Using Birdsall et al's (2000) definition would exclude at least 25% of the middle class at each end of the current distribution

Robustness Checks

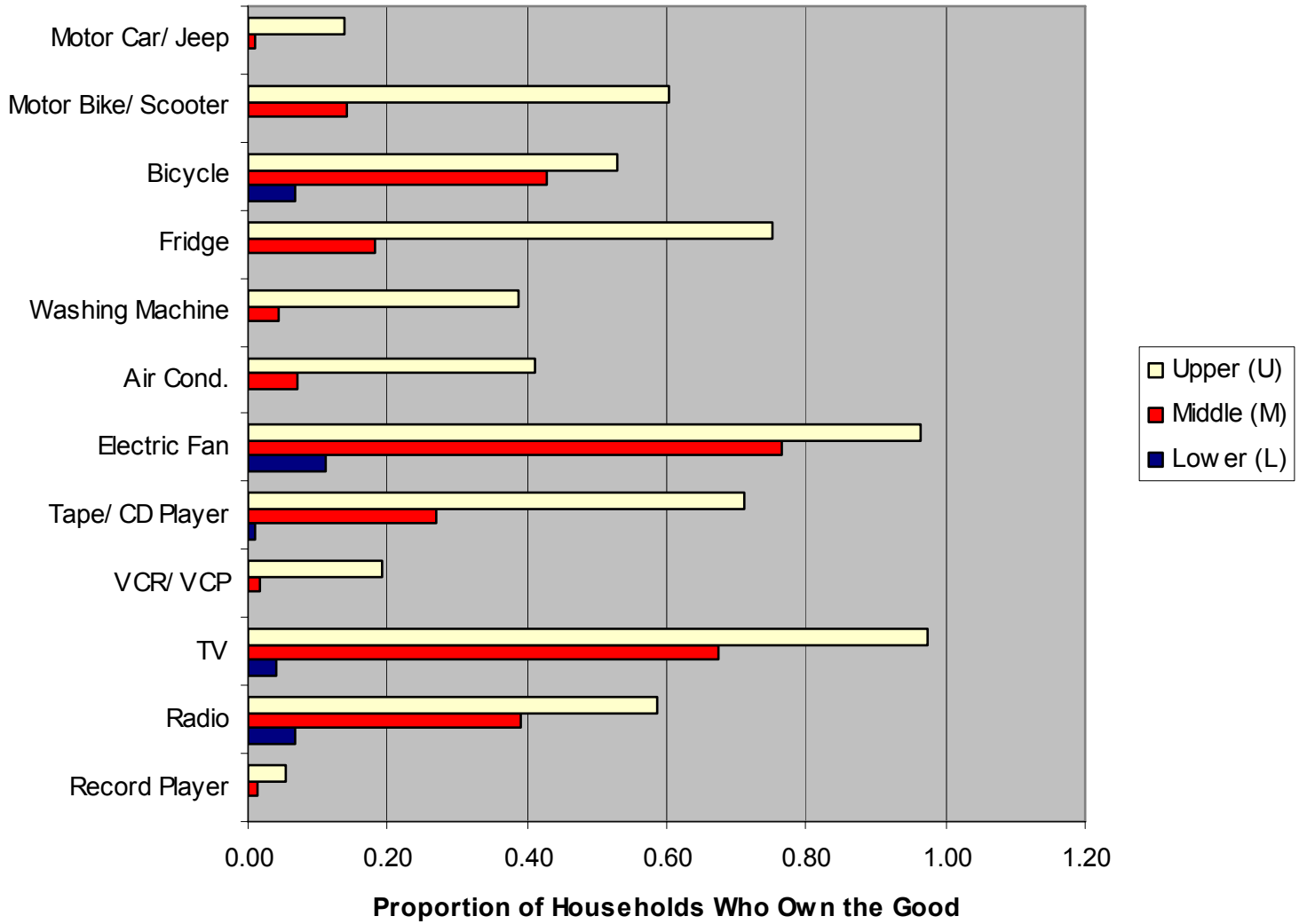
- Control for age: Results unchanged
- Add more goods: Results caution against adding 'necessary' goods to the analysis

Conclusion

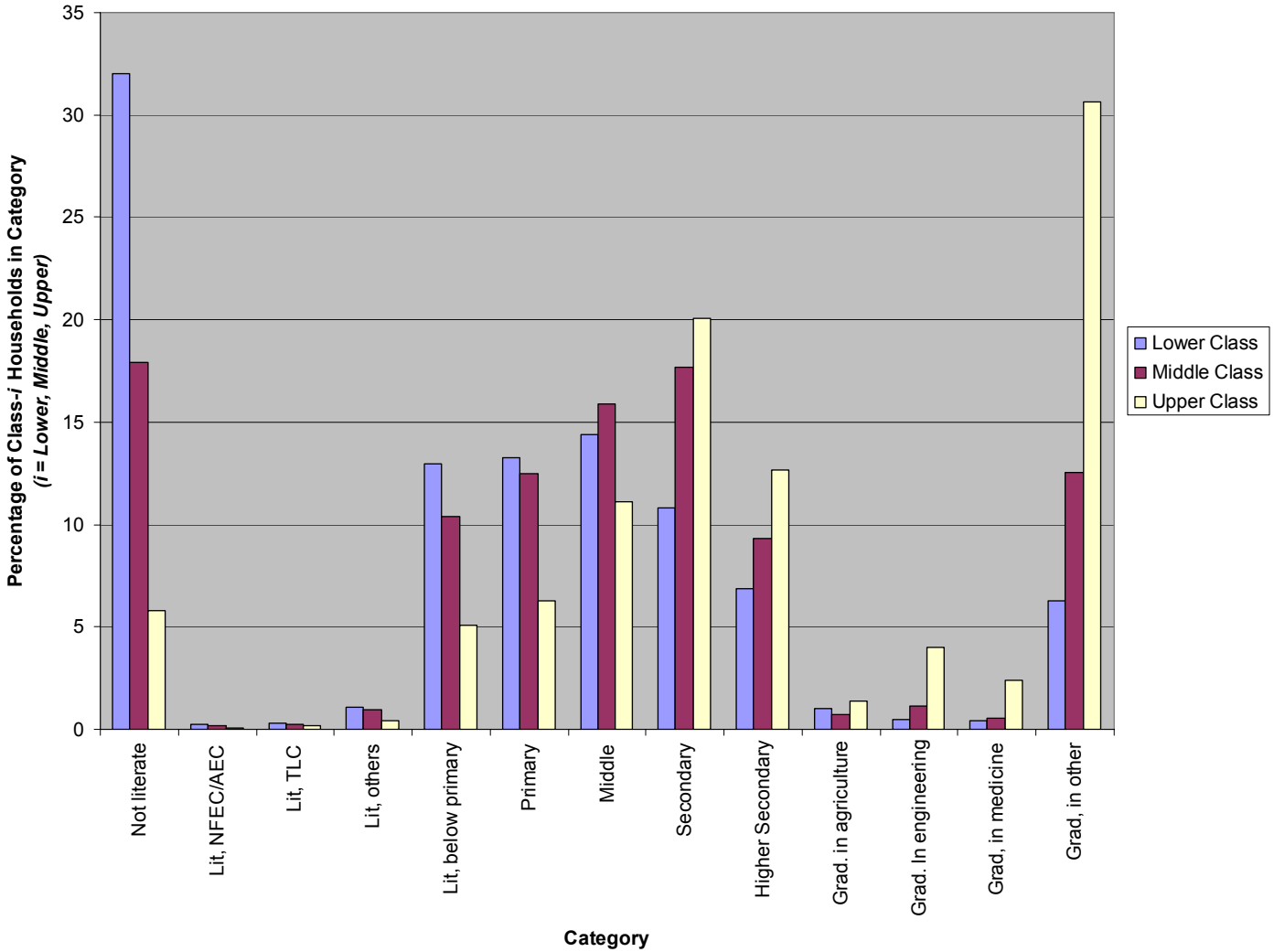
- The Mixture Model estimated by the EM algorithm provides a robust framework for identifying and estimating the size of the middle class (defined by their durable ownership habits)
  - Independent of ad hoc assumptions about who constitutes the middle class
  - Comparability of the middle class across countries

- Estimates suggest that the urban Indian middle class has lower per capita expenditure than previously assumed and is larger than previously estimated

Appendix: Ownership of Goods, by Assigned Class



Appendix: Educational Attainment of Household Head, by Assigned Class



Appendix: Robustness Checks: Control for Age

			Class Proportions (π)			Binomial probabilities (p)			Mean # goods (12p)		
Sample	Sample size	Proportion of Entire Sample	L	M	U	L	M	U	L	M	U
(1) All households	48,924	1	0.20	0.62	0.18	0.03	0.25	0.52	0.3	3.0	6.3
(2) Household head 35 years or less	14,834	0.30	0.22	0.67	0.11	0.02	0.20	0.46	0.2	2.4	5.6
(3) Household head between 36 and 65 years	30,738	0.63	0.18	0.62	0.20	0.03	0.27	0.54	0.3	3.3	6.4
(4) Household head 66 years or more	3,352	0.07	0.20	0.58	0.22	0.04	0.29	0.56	0.5	3.4	6.8
Estimates of Class Proportions and Means, from (2) - (4) →			0.19	0.63	0.17				0.3	3.0	6.2

Appendix: Robustness Checks: Add More Goods?

- NSS questionnaire has 'ownership' questions on 10 durable items not included in the analysis
  - 7 furniture & fixture items: bedstead, almirah/dressing table, chair/stool/bench/table, luggage, pillows, carpet/ durries, paintings
  - 3 household items: sewing machine, stove, pressure cooker
- The 12 items used in the EM analysis have been chosen to be consistent with the literature (Senauer & Goetz (2004)) and because these goods are indicative of affluence
- Repeat the analysis using all 22 goods for which data is available
- Results caution against adding 'necessary' goods to the analysis

Appendix: Robustness Checks: Add more goods? (All 22 goods included)

Class	Share of Urban Population (π_i)	Probability of Owning a Durable (p_i)	Mean No. of Durables Owned ($12p_i$)
Lower	11%	0.03	0.4
Middle	29%	0.37	4.5
Upper	60%	0.78	9.3

Appendix: Future Work
Preliminary Findings: Discriminating Goods

	Independent Variables in Individual Regressions											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Record Player	Radio	TV	VCR/ VCP	Tape/ CD Player	Fan	AC	Washer	Fridge	Bicycle	Motor Bike	Car/ Jeep
Dep. Var. Estimated Probability of being in:												
L	- 0.2	- 0.26	- 0.47	- 0.21	- 0.28	- 0.51	- 0.23	- 0.22	- 0.27	- 0.27	- 0.25	- 0.21
M	- 0.15	0.09	0.19	- 0.4	- 0.08	0.27	- 0.28	- 0.39	- 0.22	0.15	- 0.22	- 0.44
U	0.35	0.17	0.28	0.61	0.36	0.24	0.5	0.61	0.49	0.12	0.47	0.65

Appendix: Rural Classes

- Extension to rural sub-sample
  - 71,385 households
  - 12 goods included in analysis
- Class size as share of entire population:
  - Rural lower: 31%
  - Rural middle: 35%
  - Rural upper: 6%

Class	π	p	Mean # gds.
L	~ 43%	~ 0.03	~ 0.4
M	~ 49%	~ 0.17	~ 2
U	~ 8%	~ 0.41	~ 5

Appendix: Summary

- Urban lower: 6%
- Urban middle: 17%
- Urban upper: 5%

- Rural lower: 31%
- Rural middle: 35%
- Rural upper: 6%