Modeling of Expenditure Distribution by a Lognormal Mixture

Serguei Aivazian

Central Economics and Mathematics Institute,

Russian Academy of Sciences

Stanislav Kolenikov

University of North Carolina at Chapel Hill

Center for Economic and Financial Research, Moscow, Russia

Motivation

Deficiencies of the official methodology

- Income distribution is modeled by a lognormal distribution
- Tails underrepresented
- Discrepancies between macro and sample averages

We propose:

- Expenditure rather than income
- The distribution is a mixture of lognormal components
- Transition: changes in labor demand $\Rightarrow$ discrete mixture
- Control for sample biases wrt household wealth
- Additional unobserved stratum of super rich sub-population

Likelihood function

$$f(x \mid \theta) = \sum_{k=1}^{K} \phi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2} \right]$$

$$\theta = \left(K, \mu_1, \sigma_1^2, \phi_1, \mu_2, \sigma_2^2, \phi_2, \ldots\right), \quad \sum_{k=1}^{K} \phi_k = 1$$

Constant variance version: $\sigma_1^2 = \sigma_2^2 = \ldots$

References

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56** (3), 463–474.

Hathaway, R. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *Ann. Stat.*, **13** (2), 795–800.

Basford, K. E., G. J. McLachlan (1985). Likelikood Estimation with Normal Mixture Models. *Appl. Stat.*, **34**, 282–289.

Kiefer, J., J. Wolfowitz (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Ann. Math. Stat.*, **27**, 887–906.

McLachlan, G.J., and D. Peel (2000). *Finite mixture models*. New York, Wiley.

Difficulties

- Homo/heteroskedastic?

  Heteroskedastic: ML estimate need not exist

  Comparison? Models are non-nested…

- Number of components?

  Likelihood ratio has an unusual distribution (estimation on the boundary?)

  McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.*, **36**, 318–324.

  Feng, Z. D., C. E. McCulloch (1996). Using bootstrap likelihood ratios in finite mixture models. *J. of the Royal Stat. Soc.*, B, **58**, 609–617.

  Information criteria (AIC, SBIC, ICOMP, whatever…)

  Goodness of fit (Pearson χ^2, Kolmogorov-Smirnov?)

  Agha, M., D. S. Branker (1997). Algorithm AS 317. Maximum Likelihood Estimation and Goodness-of-fit Tests for Mixtures of Distributions. *Appl. Stat.*, **46** (3), 399--407.

Data

Russian Longitudinal Monitoring Survey

http://www.cpc.unc.edu/rlms/

Carolina Population Center and

Institute of Sociology, RAS

- Multistage clustered design
- 38 strata, of which 3 are self-representative metropolitan areas; 1 PSU per stratum
- 4718 households in the design
- 3600-3800 households (~10 ths. individuals) actually participating
- panel study
- got the refusal data from organizers

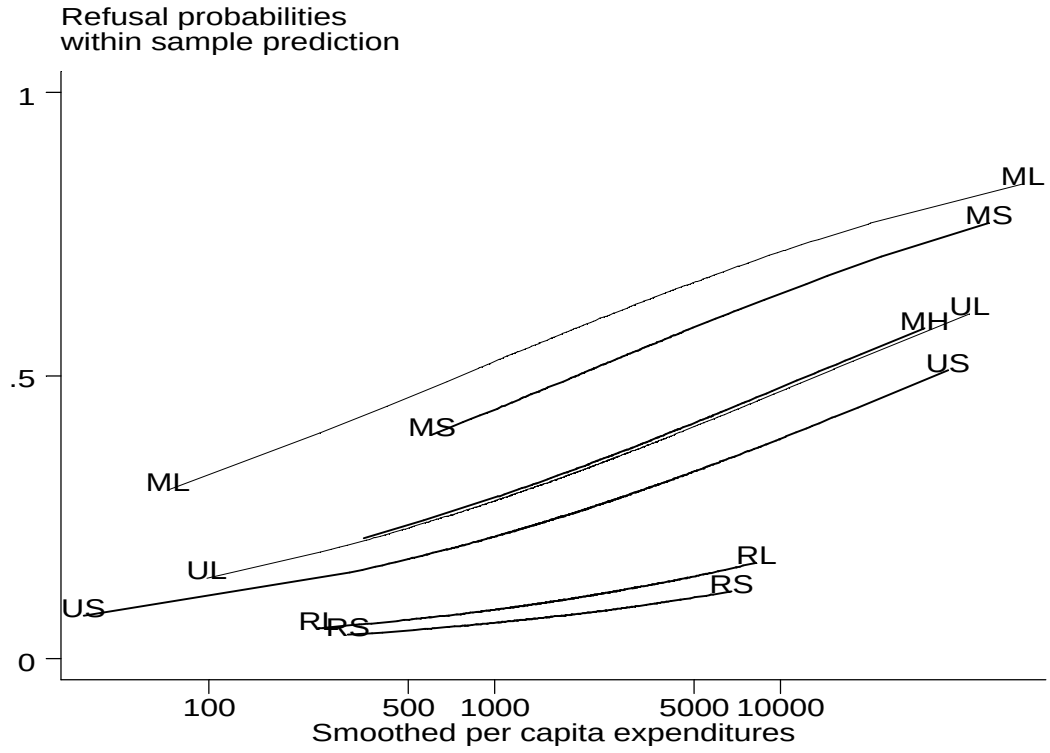
Mroz, T., L. Henderson, and B.M. Popkin (2001). "Monitoring Economic Conditions in the Russian Federation: The Russian Longitudinal Monitoring Survey 1992-2000." *Report submitted to the U.S. Agency for International Development.* Carolina Population Center, University of North Carolina at Chapel Hill.

Weights

Apart from the design weights that account for probability to be included into the sample by design, we wanted also to account for probability of non-response that we believed to be related to the household wealth.

Prob[*refrain from survey at least once*]=logit(**.**)

log(Wealth)	Urban, metro	Education
0.399 (0.079)**	++	U-shaped



Implementation and results

Stata 6 module available from author's website
http://www.komkon.org/~tacik/stata

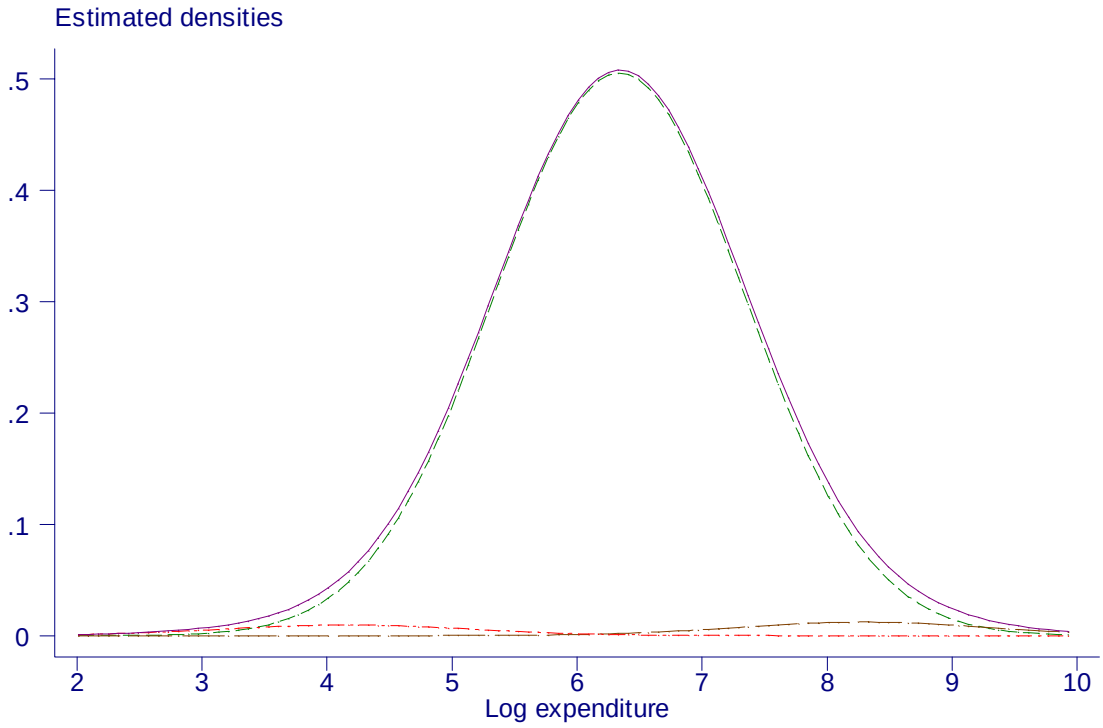
- Parameter transformations
  To ensure numerical stability and maximization without constraints: $\sigma^2 \rightarrow \log(\sigma^2)$; $\phi \rightarrow$ multinomial logit(ϕ)
- Convergence
  Declared if changes to the likelihood are small, and/or the changes in the estimated parameters are small, and/or gradient is small.
  Restart if the estimated covariance matrix is singular, or too many iterations performed.
- Multiple maxima
  Yes, there are. If the number of components is greater than the "optimal" one, then you are bound to find 3-5-… maxima.
- Bad identification
  Two or more components can stick together; most of the time diagnosed by the convergence tracer.
- Large samples curse
  Sample sizes $1\text{-}3\cdot10^3$: χ^2 statistic is U-shaped wrt K.
  Sample size 10^4: χ^2 statistic is ≥ 50.
- Outlier sensitivity
  Especially for the heteroskedastic model

Distribution estimates

The “best” models identified within the homoskedastic version included three components, with the dominant modal one.

Mean log expenditure	Share of the component
6.340	95.8%
8.282	2.3%
4.159	1.8%

The estimate of sigma = 0.756

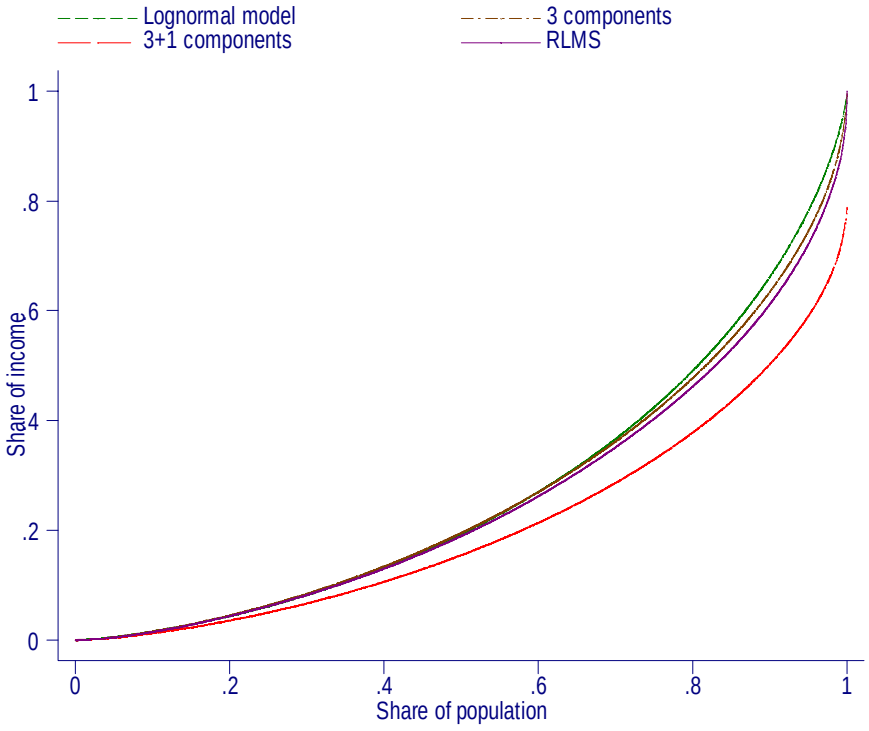


Unobserved households

Assumption: there is a completely unobserved sub-population at the top end of the distribution

- expenditure well above those in the sample
- responsible for the discrepancy between the macro and sample averages (1211 vs 932 rubles)
- modeled as an additional lognormal component with a very small population share

When the component is included, Lorenz curve shifts down dramatically, and Gini increases from 0.48 to 0.61.



Sensitivity analysis shows gradual moves of the Lorenz curve and changes in Gini value, respectively.

Conclusions

- Lognormal approximation is inadequate but not awful
- Mixture model is a better approximation, although rather computationally intensive
- Wealthy population underrepresented, arguably quite seriously
- Parametric bootstrap reconstructs the unobserved stratum: Gini 0.5-0.6.

Published as:

Aivazian, S., and S. Kolenikov (2001). Poverty and Expenditure Differentiation of Russian Population. *EERC Working Paper*, #01/01E.

http://www.eerc.ru/publications/workpapers/WP_01-01e.pdf

Further development (?)

- EM algorithm: update means & variances — update proportions

Peters, B. C. Jr., H. F. Walker (1978). An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM J. of Appl. Math.*, **35** (2), 362–378.

Xu, L., M. I. Jordan (1996). On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, **8**, 129–151.

- penalized likelihood for difference in variances

Large sample

weighted noinit: 9176 observations

Results with strata <1% are discarded (most of the runs with 4-5 components)

#	LL, +11000	Chi2	df	p	AIC, -23000	SBIC, -23000	sigma	m1	share 1	m2	share 2	m3	share 3	m4	share 4	m5	share 5
1 component																	
20	-684.61	175.18	11	0.00	387.46	.	.865	6.343									
2 components																	
9	-618.34	124.08	9	0.00	244.67	273.17	.826	6.370	.989	3.914	.011						
10	-633.65	125.11	9	0.00	275.29	303.79	.838	6.326	.9936	8.968	.0064						
Model 1 identified: 5																	
3 components																	
18	-532.21	76.867	7	0.00	76.42	119.17	.756	6.340	.958	8.282	.023	4.159	.018				
Model 2.2 identified: 3																	
Model 2.1 identified: 2																	
5 components																	
8	-515.97	69.32	3	0.00	51.95	123.19	.684	6.294	.8762	3.022	.0022	9.766	.0023	4.652	.0354	7.562	.0840
Model 3.1 identified: 5																	
Model 2.2 identified: 3																	

Smaller sample

3619 observations (1 outlier)

	initial	improve	Iteration 0	Last iteration	AIC	ICOMP	Chi2()	Prob	Freq	Components
1				(1)	3564.37	3565.68	65.55 (12)	0.000	always	Mode
2	-2727.33	-2093.25	-2000.10	-1730.96 (5/9)	3461.93	3473.32	43.27 (10)	0.000	5	Mode + R hump (<1%)
2	-2727.33	-1896.25	-1896.25	-1761.27 (6/9)	3522.53	3534.15	43.80 (10)	0.000	4	Mode + L hump (1.1%)
2									3	Mode^2
3	-2789.54	-1810.20	-1810.20	-1698.39 (6/7)	3396.78	3410.52	17.40 (8)	**0.026**	5	Mode + L hump (2.0%) + R hump (0.5%)
3	-2789.54	-1799.46	-1729.60	-1710.85 (8)	3421.71	3439.75	33.88 (8)	0.000	1	Mode + R hump (1.5%)+ outlier
3				-1730.96 -1761.27					4	Mode + hump^2 or Mode^2 + hump
4				-1698.39					6	Mode^2 + L hump (2.0%) + R hump (0.5%)
4	-2814.77	-2482.82 -2170.17	-1830.13 -2066.80	-1669.55 (6/24)	3339.09	3367.64	12.00 (6)	**0.062**	3	Mode + L hump (2.9%) + R hump (1.4%) + outlier
4	-2826.54	-2043.54	-1777.81	-1698.22 (8)	3396.44	3482.75	17.42 (6)	0.008	1	Mode + L hump (2.1%) + LL hump (<.1%) + R hump (1.1%)
5	-2826.54	-2052.48 -1997.30 -1752.56	-1794.78 –1707.12 -1697.35	-1668.41 (7/8)	3336.82	3361.98	11.04 (4)	**0.026**	3	Mode + L hump (3.6%) + LL hump (0.1%) + R hump (1.4%) + outlier
5	-2826.54	-1912.52 -1838.38	-1900.05 -1687.89	-1667.17 (8/16)	3334.34	3361.70	9.22 (4)	**0.056**	2	Mode + R hump (3.1%) + L hump (3.2%) + RR hump (0.3%) + outlier
5				-1669.55					4	Mode^2 + L hump (2.9%) + R hump (1.4%) + outlier