Structural Change and Economic Dynamics 23 (2012) 69–91

Contents lists available at SciVerse ScienceDirect

Structural Change and Economic Dynamics

journal homepage: www.elsevier.com/locate/sced

The distribution of household consumption-expenditure budget shares

Matteo Barigozzi[a], Lucia Alessi[b], Marco Capasso[c], Giorgio Fagiolo[d,*]

[a] London School of Economics, London, UK
[b] European Central Bank, Frankfurt am Main, Germany
[c] UNU-MERIT and School of Business and Economics, Maastricht University, The Netherlands
[d] Sant'Anna School of Advanced Studies, Laboratory of Economics and Management, Pisa, Italy

ARTICLE INFO

Article history:
Received 11 February 2011
Received in revised form
26 September 2011
Accepted 27 September 2011
Available online xxx

JEL classification:
D3
D12
C12

Keywords:
Household consumption expenditure
Budget shares
Sum of log-normal distributions

ABSTRACT

This paper explores the statistical properties of household consumption-expenditure budget share distributions – defined as the share of household total expenditure spent for purchasing a specific category of commodities – for a large sample of Italian households in the period 1989–2004. We find that household budget share distributions are fairly stable over time for each specific category, but profoundly heterogeneous across commodity categories. We then derive a parametric density that is able to satisfactorily characterize (from a univariate perspective) household budget share distributions and: (i) is consistent with the observed statistical properties of the underlying levels of household consumption-expenditure distributions; (ii) can accommodate the observed across-category heterogeneity in household budget-share distributions. Finally, we taxonomize commodity categories according to the estimated parameters of the proposed density. We show that the resulting classification is consistent with the traditional economic scheme that labels commodities as necessary, luxury or inferior.

1. Introduction

The study of household budget allocation – i.e., how the budget of a household is allocated to buy different commodities – is one of the most traditional topics in economics (Prais and Houthakker, 1955). Household budget shares contain useful information to shed light on this issue. Indeed, the household budget share for a given commodity category g is defined as the ratio between the expenditure for the commodity category g and total household resources, as measured by, e.g., total expenditure or total income.

In the last decades, this topic has received a lot of attention by applied economists. In particular, many efforts have been devoted to develop statistical demand functions for homogeneous groups of commodities, e.g., by relating the expenditure of consumers or households for a given commodity category to prices and individual-specific variables as total expenditure or income, household size, head-of-household age, and so on (Deaton, 1992; Blundell, 1988). A paradigmatic example in this line of research is the analysis of Engel curves (Engel, 1857; Deaton, 1992; Blundell, 1988; Chai and Moneta, 2010), which describe how the expenditure for a given commodity category varies with household's total resources (Lewbel, 2008).

Such a research program has been mostly characterized by a theory-driven approach (Attanasio, 1999). In fact, the parametric specifications that are employed in the estimation of each specific demand function are in general taken to be consistent (albeit in a weak way) with some underlying theory of household expenditure behavior, which very often is the standard model based on utility maximization undertaken by fully rational agents (Banks et al., 1997; Blundell et al., 2007). Furthermore, no matter whether parametric or non-parametric techniques are employed, the estimation of demand systems

* Corresponding author.
E-mail address: giorgio.fagiolo@sssup.it (G. Fagiolo).

0954-349X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.strueco.2011.09.003

or Engel curves compresses household heterogeneity – for any given income or total expenditure level – to the knowledge of the first two moments (at best) of household expenditure level or budget share distribution for the commodity category under study.

This of course is fully legitimate if the aim of the researcher is to empirically validate a given theoretical model, or if there are good reasons to believe that the distribution under analysis can be fully characterized by its first two moments. However, from a more data-driven perspective, constraining in this way the exploration of the statistical properties of the observed household expenditure patterns may be problematic for a number of reasons.

First, heterogeneity of household consumption-expenditure patterns is widely considered as a crucial feature because, as Pasinetti (1981) notices: "At any given level of per capita income and at any given price structure, the proportion of income spent by each consumer on any specific commodity may be very different from one commodity to another". This suggests that, in order to fully characterize such heterogeneity, one should perform distributional analyses that carefully investigate how the shape – and not only the first two moments – of household consumption expenditure and household budget share distributions change over time and between different commodity categories.

Second, understanding heterogeneity may be important to build sound micro-founded, macroeconomic, consumption models that go beyond the often disputable representative-agent assumption (Kirman, 1992; Hartley, 1997; Gallegati and Kirman, 1999). For example, Caselli and Ventura (2000) show that models based on the representative-agent assumption impose almost no restrictions on household consumption expenditure and budget share distributions. On the contrary, Forni and Lippi (1997) demonstrate that heterogeneity is crucial when aggregating individual behavior in macro models. Furthermore, Ibragimov (2005) provides support to the insight that higher-than-two moments can have a relevant impact on the dynamics of macro models (on these and related points, see Hildenbrand, 1994, among others).

Third, adopting a more theory-free approach focused on distributional analysis may help to discover fresh stylized facts related to how households allocate their consumption expenditures across different commodity categories. In fact, theory-free approaches aimed at searching for stylized facts are not new in economics and econometrics (see *inter alia* Kaldor, 1961; Hendry, 2000). More recently, this perspective has been revived in the field of econophysics, where the statistical properties of many interesting micro and macro economic variables (e.g., firm size and growth rates, industry and country growth rates, wealth and personal income) have been successfully characterized by using parametric techniques (Chatterjee et al., 2005; Clementi and Gallegati, 2005; Axtell, 2001; Bottazzi and Secchi, 2006; Fagiolo et al., 2008). These studies show that, despite the turbulence typically detected at the microeconomic level (e.g., entry and exit of firms; positive and negative persistent shocks to personal income), there exists an incredible high level of regularity in the shape of microeconomic cross-section distributions, both across years and countries.

Notwithstanding such successful results, similar distributional analyses have not been extensively performed, so far, on consumption-related microeconomic variables such as household consumption expenditures and budget shares, for which reliable and detailed cross-section data are also available. This is somewhat surprising because – as Attanasio (1999) notices – understanding consumption is crucial to both micro- and macro-economists, as it accounts for about two-thirds of GDP and it decisively determines (and measures) social welfare.

There are only three exceptions – to the best of our knowledge – to this lack of distributional studies on household consumption indicators. Chronologically, the first instance is the pioneering work of Aitchison (1986) on compositional data analysis. There, budget-share data are only to illustrate applications of a multivariate approach that aims at parametrically modeling data defined by construction on the simplex (see also McLaren et al., 1995; Fry et al., 1996, for an application of compositional-data analysis to consumption budget shares within the context of modified almost-ideal demand systems). Nothing is said, however, on the economic implications of such an approach, neither the method is applied to other databases. We shall go back to these points below.

More recently, Battistin et al. (2007) employ expenditure and income data from U.K. and U.S. surveys and show that total household consumption expenditure distributions are well-approximated by log-normal densities (or, as they put it, are "more log-normal than income"). In a complementary paper, Fagiolo et al. (2010) argue that log-normality is valid only as a first approximation for Italian total household consumption expenditure distributions, while a refined analysis reveals asymmetric departures from log-normality in the tails of the distributions. Both contributions focus on characterizing, from a univariate perspective, the dynamics of household consumption-expenditure aggregate distributions only. The issue of exploring the statistical properties of household consumption expenditure or budget share distributions disaggregated among commodity categories is not addressed.

This paper is a preliminary attempt to fill these gaps. To do so, we employ data from the "Survey of Household Income and Wealth" (SHIW) provided by the Bank of Italy to study household consumption expenditure and budget share distributions for a sequence of 8 waves between 1989 and 2004. To fully exploit the database, we focus on four commodity categories: nondurable goods, food, durable goods, and insurance premia. Note that food is actually a subcategory of nondurable goods, but for its intrinsic importance we consider it as a separate commodity category throughout the paper. Since insurance premia are rarely studied in the consumption literature as a category on its own, we explicitly consider them in this study to understand whether they exhibit different distributional properties as compared to more traditional consumption categories.

We aim at empirically investigating the statistical properties of *unconditional* household budget share

distributions (and consumption expenditure distributions) of these four commodity categories and their dynamics with a parametric approach, where by *unconditional* distributions we mean here not conditioned to total household resources, i.e., income or total expenditures. More specifically, we look for a unique, parsimonious, closed-form univariate density family that: (i) is able to satisfactorily fit observed unconditional household budget share distributions, so as to accommodate the existing heterogeneity emerging across households, among different commodity categories and over time; (ii) is consistent with the statistical properties of the (observed) household consumption expenditures distributions employed to compute budget share distributions; (iii) features economically interpretable parameters that, once estimated, can help one to build economically meaningful taxonomies of commodity categories.

We begin with a descriptive analysis aimed at empirically exploring the stability of household budget share distributions over time. Estimated sample moments show that the shape of the household budget share distribution of each given commodity category does not dramatically change over the time interval considered. However, for any given wave, there emerges a lot of across-commodity heterogeneity in the observed shapes of household budget share distributions. More precisely, we show that the underlying household-consumption expenditures univariate distributions – for any given wave and commodity category – are well-proxied by log-normal distributions (with very different parameters).

Yet, a preliminary multivariate investigation indicates that existing multivariate parametric density families defined on the simplex (i.e., describing shares that sum to one; see Aitchison and Egozcue, 2005, for a review) are not able to satisfactorily model our data. Therefore, we turn to a univariate approach and we derive an original family of densities, defined over the unit interval, which is consistent with the detected log-normality of household consumption expenditures distributions. The precise formulation of the closed-form density can be shown to depend on the chosen approximation for the random variable defined as the sum of (possibly correlated) log-normal distributions. In the literature there exist two possible approximations, namely the log-normal and theinverse-Gamma, which we both fit to our data. To benchmark our results, we also fit household budget share distributions with univariate Beta distributions, which are in principle very flexible densities defined over the unit interval but lack any consistency with the shape of the random variables which household budget shares stem from.

We find that, in the case of Italy, for all the waves under study and for all the commodity categories, the proposed density family – using either approximation – outperforms the Beta in fitting observed household budget share distributions for the majority of cases. Indeed, according to simple measures of goodness-of-fit (e.g., the average absolute deviation), the proposed density family is able to better accommodate the existing shape-heterogeneity that characterizes household budget share distributions across different commodity categories. Furthermore, the estimated parameters of the proposed density allow to reproduce an economically meaningful taxonomy of commodity categories, which interestingly maps into the traditional classification of commodities among necessary, luxury or inferior goods.

The paper is structured as follows. In Section 2 we describe the database that we employ in the analysis and we discuss some methodological issues. Section 3 presents a preliminary descriptive analysis of household consumption expenditure and budget share distributions, whereas Section 4 discusses multivariate analyses. In Section 5 we derive the proposed family of univariate theoretical densities. Section 6 presents fitting results obtained with that density family, and compares them with Beta variates. Section 7 briefly reports on some interpretations of our exercises in terms of commodity category taxonomies. Finally, Section 8 concludes.

2. Data and methodology

The empirical analysis below is based on the "Survey of Household Income and Wealth" (SHIW) provided by the Bank of Italy. The SHIW is one of the main sources of information on household income and consumption in Italy. Indeed, the quality of the SHIW is nowadays very similar to that of surveys in other countries like France, Germany and the U.K. SHIW data are regularly published in the Bank's supplements to the Statistical Bulletin and made publicly available online[1] (see Brandolini, 1999; Battistin et al., 2003 for additional details).

The SHIW was firstly carried out in the 1960s with the goal of gathering data on income and savings of Italian households. Over the years, the survey has been widening its scope. Households are now asked to provide, in addition to income and wealth information, also details on their consumption behavior and even their preferred payment methods. Since then, the SHIW was conducted yearly until 1987 (except for 1985) and every two years thereafter (the survey for 1997 was shifted to 1998).

The present analysis focuses on the period 1989–2004. We therefore have 8 waves. The sample used in the most recent surveys comprises about 8000 households (about 24,000 individuals distributed across about 300 Italian municipalities). The sample is representative of the Italian population and is based on a rotating panel targeted at 4000 units. Available information includes data on household demographics (e.g., age of household head, number of household components and geographical area), disposable income, consumption expenditures, savings, and wealth.[2]

In this study, we employ yearly data on (nominal) aggregate, household consumption expenditures and on the following disaggregated commodity categories: nondurable goods (N), durable goods (D), and insurance premia (I). Nondurable goods include also food (F), which we

[1] http://www.bancaditalia.it/statistiche/indcamp/bilfait.

[2] More precisely, the sample unit is the household defined in a broad sense as a group of individuals related to each other by links of blood, marriage or affection, living together and pooling (part of) their incomes. The survey was restricted to households with at least two members. Data were not aggregated across family sizes. For more details, see Brandolini (1999).

consider as a separate (sub-)category of commodities. According to the definition of the Bank of Italy, expenditures for nondurable goods correspond to all spending on both food and non-food items, excluding expenses for durable goods and insurance, maintenance, mortgage and rent payments. The expenditures for food include spending on food products in shops and supermarkets, and spending on meals eaten regularly outside home. Household expenditures for durable goods correspond to items belonging to the following categories: precious objects, means of transport, furniture, furnishings, household appliances, and sundry articles. Finally, the commodity category labeled as "insurances" includes the following forms of insurance: life insurance, private or supplementary pensions, annuities and other forms of insurance-based saving, casualty insurance, and health insurance policies. Note that insurance expenditure, *strictu sensu*, might be considered as a form of saving. However, we consider insurances as a commodity category for basically two reasons: (i) insurance forms might also be seen as consumption goods, insomuch as they cover actual expenses which are borne by the household (e.g., for pharmaceutical products); (ii) the insurance itself might be considered – and indeed appears in theoretical models – as a good, which an agent might or might not purchase: the only difference with respect to a traditional consumption good stands in the fact that the degree of risk aversion influences the amount of insurance purchased. The major drawback of the SHIW database is that it does not allow for further disaggregation of consumption categories into more detailed groups. Furthermore, the disaggregation level of categories as durable goods and insurance premia is not totally comparable, as the former certainly comprises much more items than the latter. As already discussed above, however, insurance premia are not frequently studied in the consumption literature as a category on its own. Therefore, it seems interesting to explicitly address here its study to better understand the extent to which its statistical properties differ from those of more traditional consumption categories.

The SHIW database includes a variable recording household total (aggregate) expenditure. This quantity is reported by households in the SHIW independently on their expenditure for disaggregated commodity categories. Therefore, it does not necessarily correspond to the sum of expenditures of the three macro consumption categories considered here (N, D and I), the sum making up on average 80% of total expenditures. In what follows, we shall employ total household expenditure as a proxy for total household resources (more on that below). Household budget shares are accordingly computed as ratios of nominal yearly quantities. More formally, our data structure consists of the distribution of yearly household budget shares defined as:

$$B_i^{h,t} = \frac{C_i^{h,t}}{C^{h,t}}, \tag{1}$$

where $t \in T = \{1989, 1991, 1993, 1995, 1998, 2000, 2002, 2004\}$ are survey waves, $i \in \{\mathrm{N}, \mathrm{F}, \mathrm{D}, \mathrm{I}\}$ are the four commodity categories, $C_i^{h,t}$ is the (nominal) consumption expenditure of household $h = 1, \ldots, H_t$ for the commodity category i, and $C^{h,t}$ is the (nominal) total consumption expenditure of household h, as reported in the SHIW. All household consumption expenditure observations have been preliminary weighted using appropriate sample weights provided by the Bank of Italy. Weighting ensures that socio-demographic marginal distributions are in line with the corresponding distributions found in Italian Statistical Office (ISTAT) population statistics and labor force surveys.[3] Outliers – defined as observations greater than 10 standard deviations from the mean – have been removed.[4] Since in each wave there were some cases of unrealistic (e.g., zero or negative) aggregate consumption expenditure figures, we dropped such observations and we kept only strictly positive ones. We also dropped households for which yearly expenditures for at least one commodity was larger or equal to total expenditure (as reported in the SHIW). Since we rule out borrowing, $B_i^{h,t} \in (0, 1)$.

Finally, we excluded zero observations from the analysis, for the following reasons: (i) it is not clear whether zero entries for nondurable goods mean a null consumption expenditure or rather they are due to mistakes in data collection; (ii) the decision whether to buy a durable good or an insurance or not (depending for example on whether household income exceeds some threshold) is different from and precedes the decision on the budget share possibly allocated to this good; we focus on the second step of the decisional process; (iii) the degree of bunching at zero one typically finds in the distributions of durable goods and insurance expenditures is influenced by factors which are likely to vary over the business cycle; (iv) sample moments computed including zeroes would be poorly informative, given the relatively large proportion of zero observations. Therefore, we ended up with a changing (but still very large) number of households in each wave H_t (see Table 3).

In what follows, we shall also make use of an alternative representation of our data that allows one to look at budget-share vectors as points in the simplex. For each household h and wave t, consider the vector

$$\underline{B}^{h,t} = (B_{\mathrm{N}}^{h,t}, B_{\mathrm{D}}^{h,t}, B_{\mathrm{I}}^{h,t}, B_{\mathrm{O}}^{h,t}), \tag{2}$$

where $B_{\mathrm{O}}^{h,t}$ is the budget shares for all other commodity categories defined as:

$$B_{\mathrm{O}}^{h,t} = 1 - (B_{\mathrm{N}}^{h,t} + B_{\mathrm{D}}^{h,t} + B_{\mathrm{I}}^{h,t}), \tag{3}$$

Obviously this implies that $\underline{B}^{h,t}$ lies in the simplex, which in our case has a 3-dimensional Euclidean space structure. In other words, we shall define the household consumption expenditure category "Others" (O) as $C_{\mathrm{O}}^{h,t} = C^{h,t} - (C_{\mathrm{N}}^{h,t} + C_{\mathrm{D}}^{h,t} + C_{\mathrm{I}}^{h,t})$ and compute budget shares as in (1) for $i \in \{\mathrm{N}, \mathrm{D}, \mathrm{I}, \mathrm{O}\}$.

[3] An interesting and open issue concerns the impact that weighting, through population trends, may have on the distributional properties of budget shares. Additional data on phenomena such as population ageing, increasing share of foreign population, lowering fertility rate and intra-country migration flows may shed some light on the implications of demographic trends for our results.

[4] The choice of 10 standard deviations was done in order to maximize the number of observations to be retained in the sample. Our exercises show that choosing a smaller number of standard deviations does not dramatically affect the results.

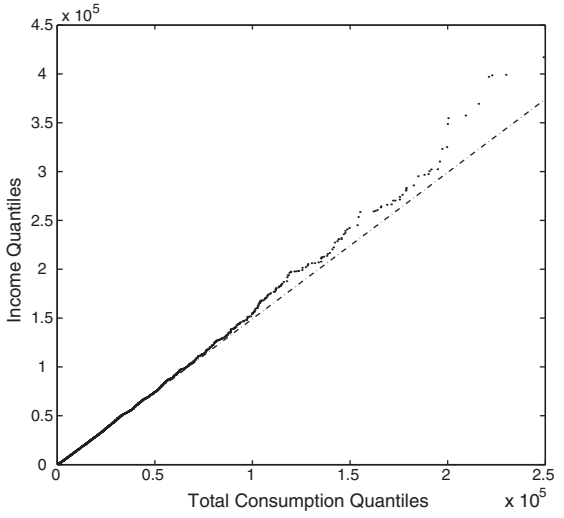

Fig. 1. Quantile–quantile plot comparing income and total consumption expenditure distributions. Wave, 2004; *Y*-axis, income quantiles; *X*-axis, consumption quantiles.

Two important points deserve to be discussed. First, we use total expenditures instead of income to proxy household total resources and compute budget shares. This is primarily done in order to separate the problem of allocating total consumption to various commodities from the decision of how much to save out of current income. Notice that this is common practice in the relevant literature. Indeed, due to the relatively higher reliability of expenditure data (as compared to income ones), most of empirical studies typically use household consumption expenditures even if theoretical models are originally developed in terms of total income (see, e.g., Banks et al., 1997). Since income is available in the SHIW database, we replicated our exercises by defining household budget shares in terms of household-income ratios without any appreciable differences in the results as far as descriptive analyses were concerned. Quantile–quantile plots comparing household income and total consumption expenditure distributions, reported in Fig. 1 for wave 2004, show that as income increases the proportion of income devoted to total consumption also increases.

Second, as already mentioned, this study is not explicitly concerned with the estimation of Engel curves, either with parametric or non-parametric approaches (Engel and Kneip, 1996; Chai and Moneta, 2008). Conversely, we treat household budget shares as agnostic variables that have an economic meaning 'per se'. Moreover, note that Engel curves describe the relationship between conditional averages of household consumption expenditures (or budget shares) for a particular commodity category and levels of income or total consumption expenditure, where averages are computed conditional to levels of income or total consumption expenditure, and possibly other explanatory variables. In this paper, we begin instead to study the statistical properties of unconditional household budget share distributions, that is – for any commodity category and wave – we pool together households irrespective of their income or total consumption expenditure, and we consequently study the shape of the ensuing distributions and their dynamics. In other words, we do not compress the overall across-household heterogeneity existing for each commodity category and wave, as done in Engel-curve studies. This is because the goal of the paper is simply to characterize the distributional shape of *unconditional* household budget share distributions and not how they change with household total budget. As we briefly recall in the concluding section, this might envisage a possible extension of the present work. One might indeed condition household budget share distributions, for each commodity category and wave, to total household resources and investigate how conditional household budget share distributions change as income or total consumption expenditure increases.

3. A preliminary statistical analysis

In this section, we begin with a preliminary, mainly descriptive, statistical analysis of Italian household consumption expenditure and budget share distributions, mainly focused on investigating whether such distributions – and their correlation structure – exhibit structural changes over time.

3.1. Household expenditure distributions

Let us start with household consumption expenditure distributions. Fig. 2 shows kernel density estimates of the logs of household consumption expenditure distributions for waves 1989, 1993, 1998, and 2002 (for the sake of exposition).[5] A preliminary visual inspection of the four panels indicates that, with the exception of insurance premia, the shape of any given household consumption expenditure distribution does not dramatically change over time. This evidence seems to be confirmed by Table 1, where we report estimated sample moments for logged household consumption expenditure distributions, and by Fig. 3, which shows their evolution over time. We will go back to this question in more details in the next section. For the moment, notice that sample means show a positive trend in time because we are considering nominal quantities, but we expect also real values to display a rightward average shift due to economic growth. However, insurance expenditure distributions display a more pronounced trend, which is probably due to the observed structural increase in expenditure for insurance premia from the late 90s also in real terms. Furthermore, Table 2 shows sample correlations and *p*-values for the null hypothesis of no correlation

[5] All kernel density plots in the paper have been obtained using density estimates computed with the function *kdens* from the Stata software package (http://ideas.repec.org/c/boc/bocode/s456410.html). Densities have been estimated on 100 points for the log consumption levels, and 50 points for the budget shares. A Gaussian kernel has been used. The bandwidth has been selected according to the "rule of thumb" suggested by Silverman (1986). Although several refinements of the "rule of thumb" have been presented in the literature (see, e.g., Sheather and Jones, 1991), we opted for the original method by (Silverman, 1986) because it was the only one minimizing the impact of outlying bins.

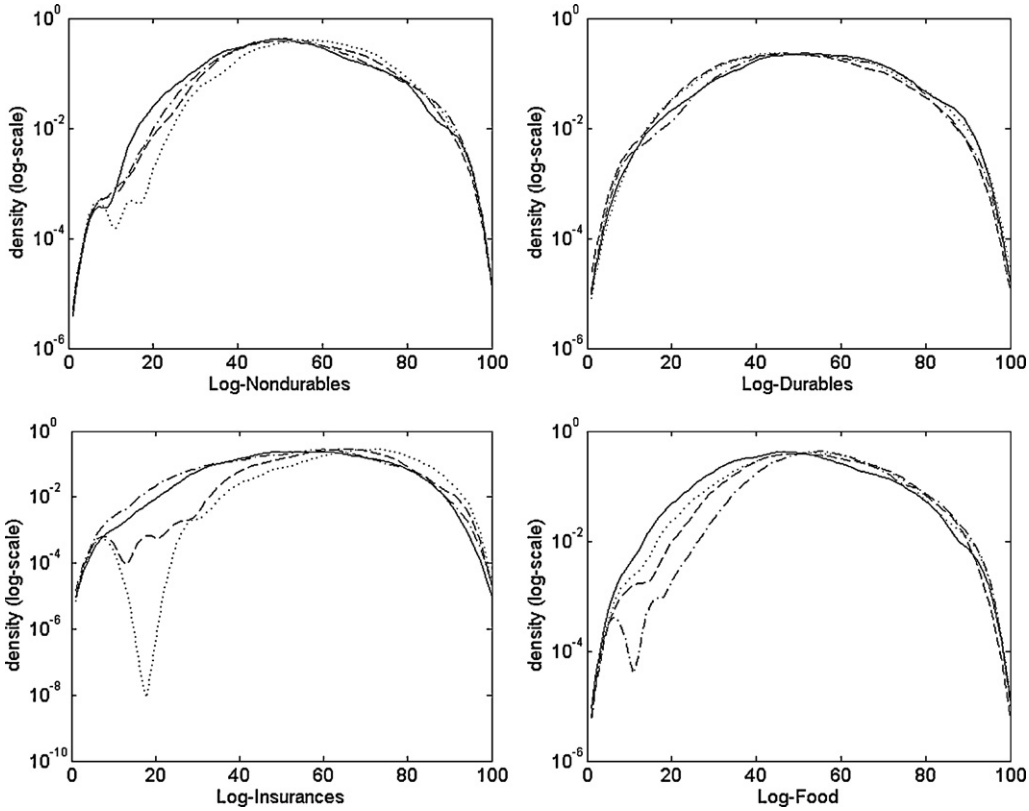

Fig. 2. Kernel-density estimates of logged household consumption expenditure distributions: evolution over time. Solid line, year 1989; dashed-dotted line, year 1993; dotted line, year 1998; dashed line, year 2002.

between household consumption expenditure distributions for different commodities in 2004, where correlation coefficients are computed (here and in what follows) only for households with non-zero expenditure for all commodity categories. As expected, the correlations among household consumption expenditure distributions are all strongly positive and significant.

The parabolic shape of logged household consumption expenditure kernels in Fig. 2 hints to the possibility that the (disaggregated) distributions might be well-proxied by log-normal densities. Notice that the existing literature shows that *aggregate* household consumption expenditure distributions are typically log-normally distributed (Battistin et al., 2007). In Fagiolo et al. (2010) we show that, as a first approximation, similar evidence is true also for the Italian total household consumption expenditure. It is then worthwhile to check if also commodity-disaggregated household consumption expenditures behave the same way. Fig. 4 indicates that a log-normal density provides reasonable fits also for our household consumption expenditure distributions disaggregated across our commodity categories. Table 1 confirms this finding, as the logs of disaggregated household consumption expenditure distributions exhibit skewness and kurtosis values very close to what would be expected if the original distributions were log-normal (i.e., 0 and 3 respectively).

As discussed in Battistin et al. (2007), consumption and income data generally suffer from under reporting (especially in the tails) and outliers, and Italian data are not an exception (Brandolini, 1999). In order to minimize the effect of gross errors and outliers, we have employed robust statistics to estimate the moments of household consumption expenditure distributions (Huber, 1981). More specifically, we have estimated the third moment with quartile skewness (Groeneveld and Meeden, 1984) and kurtosis using Moors's octile-based robust estimator (Moors, 1988). Robust-estimator analysis supports log-normality of household consumption expenditure distributions. In fact, according to standard bootstrap tests, robust skewness and kurtosis of logged household consumption expenditure distributions are often close to their expected values in normal samples (0 and 1.233, respectively). Since log-normality seems to be a good proxy also for $C_{\mathrm{O}}^{h,t}$ distributions, it is tempting to check for multivariate log-normality of the 4-dimensional vector $\underline{C}^{h,t} = (C_{\mathrm{N}}^{h,t}, C_{\mathrm{D}}^{h,t}, C_{\mathrm{I}}^{h,t}, C_{\mathrm{O}}^{h,t})$. Unfortunately, a standard Energy-test for multivariate normality (Szekely and Rizzo, 2005) strongly rejects the null hypothesis with p-value close to zero (we shall go back to this point in Section 4).

3.2. Household budget-share distributions

We turn now to a descriptive analysis of household budget share distributions. Fig. 5 shows the plots of kernel-density estimates for 1989, 1993, 1998, and 2002. Again, a

Table 1
Moments of logged household consumption expenditure distributions vs. waves. Avg, average values over the whole period; TC, total consumption; N, nondurables; D, durables; I, insurances; F, food. The figures labeled as N + D + I only refer to households with non-zero expenditure for each commodity category.

	Stats	Waves								Avg
		1989	1991	1993	1995	1998	2000	2002	2004	
N	*N* obs.	7409	7209	6223	6258	5588	6277	6361	6281	6451
	Mean	8.551	8.518	8.715	8.876	8.863	8.949	8.942	9.073	8.811
	Std. dev.	1.014	1.064	0.962	0.918	0.939	0.939	0.982	0.909	0.966
	Skewness	0.201	0.111	0.281	0.108	0.007	0.102	0.174	0.143	0.141
	Kurtosis	2.844	2.893	2.927	2.703	2.710	2.727	2.647	2.810	2.783
D	*N* obs.	2534	2352	2082	1856	2091	1920	1833	1961	2079
	Mean	6.554	6.713	6.529	6.900	6.902	7.078	6.881	6.879	6.805
	Std. dev.	1.626	1.656	1.615	1.593	1.560	1.588	1.635	1.568	1.605
	Skewness	−0.041	0.033	0.028	0.017	0.123	0.047	0.158	0.296	0.083
	Kurtosis	2.698	2.678	2.561	2.509	2.503	2.454	2.560	2.683	2.581
I	*N* obs.	1780	1928	2257	2961	2652	2575	2175	2164	2312
	Mean	5.501	5.604	5.737	5.853	6.198	6.359	6.358	6.551	6.020
	Std. dev.	1.500	1.599	1.557	1.567	1.476	1.398	1.408	1.408	1.489
	Skewness	−0.069	−0.216	−0.269	−0.284	−0.504	−0.166	−0.228	−0.279	−0.252
	Kurtosis	2.555	2.788	2.689	2.649	3.218	2.641	2.965	3.592	2.887
F	*N* obs.	7409	7228	6235	6261	5596	6281	6366	6281	6457
	Mean	7.738	7.808	8.014	8.108	8.089	8.119	8.133	8.241	8.031
	Std. dev.	0.978	1.059	0.969	0.913	0.930	0.947	0.975	0.919	0.961
	Skewness	0.214	0.109	0.253	0.099	0.017	0.108	0.138	0.116	0.132
	Kurtosis	2.753	2.844	3.029	2.744	2.770	2.805	2.635	2.853	2.804
N + D + I	*N* obs.	896	904	1099	1225	1310	1162	930	1016	1068
	Mean	9.148	9.156	9.269	9.484	9.435	9.607	9.640	9.716	9.432
	Std. dev.	0.932	1.030	0.904	0.830	0.909	0.877	0.945	0.861	0.911
	Skewness	0.326	0.078	0.335	0.186	0.011	0.174	0.232	0.109	0.182
	Kurtosis	2.848	3.131	2.818	2.534	2.718	2.420	2.558	2.721	2.718
TC	*N* obs.	7416	7237	6245	6274	5598	6282	6370	6285	6463
	Mean	8.907	8.905	9.093	9.256	9.300	9.369	9.377	9.510	9.215
	Std. dev.	1.028	1.095	0.978	0.925	0.934	0.939	0.975	0.908	0.973
	Skewness	0.230	0.150	0.240	0.120	0.077	0.132	0.262	0.236	0.181
	Kurtosis	2.828	2.865	2.855	2.690	2.659	2.704	2.659	2.866	2.766

first visual inspection does not seem to detect strong time dependence in the shape of budget-share distributions. Conversely, as expected, their shapes differ significantly across commodity categories. Household budget share distributions for nondurable goods and food are relatively bell shaped and a large mass of observations is shifted towards the right of the unit interval. Kernels of durable goods and insurance premia are instead much more right-skewed and monotonically decreasing. Note also that insurance-premia kernels exhibit a relevant irregularity in the right tail, due to a small sample-size problem. The strong across-commodity heterogeneity that clearly emerges in the shape of household budget share distributions suggests that in order to find a unique, parsimonious, parametric statistical model able to satisfactorily fit the data, one would require a very flexible density family.

Estimated sample moments of household budget share distributions are reported in Table 3. On average, 68% of total household expenditures is related to nondurable goods, while food accounts for 33% of the total. Much less is spent on durable goods and insurance premia, as they respectively represent – on average – 13% and 5% of total household consumption expenditures. No appreciable differences are found by replacing the denominator

Table 2
Correlations among household consumption expenditure distributions and *p*-values (in brackets) for the null hypothesis of no correlation. Wave 2004. TC, total consumption; N, nondurables; D, durables; I, insurances; F, food.

	N	D	I	F
D	0.40	–	–	–
	(0.00)	–	–	–
I	0.49	0.39	-	–
	(0.00)	(0.00)	–	–
F	0.87	0.29	0.44	–
	(0.00)	(0.00)	(0.00)	–
TC	0.92	0.58	0.51	0.80
	(0.00)	(0.00)	(0.00)	(0.00)

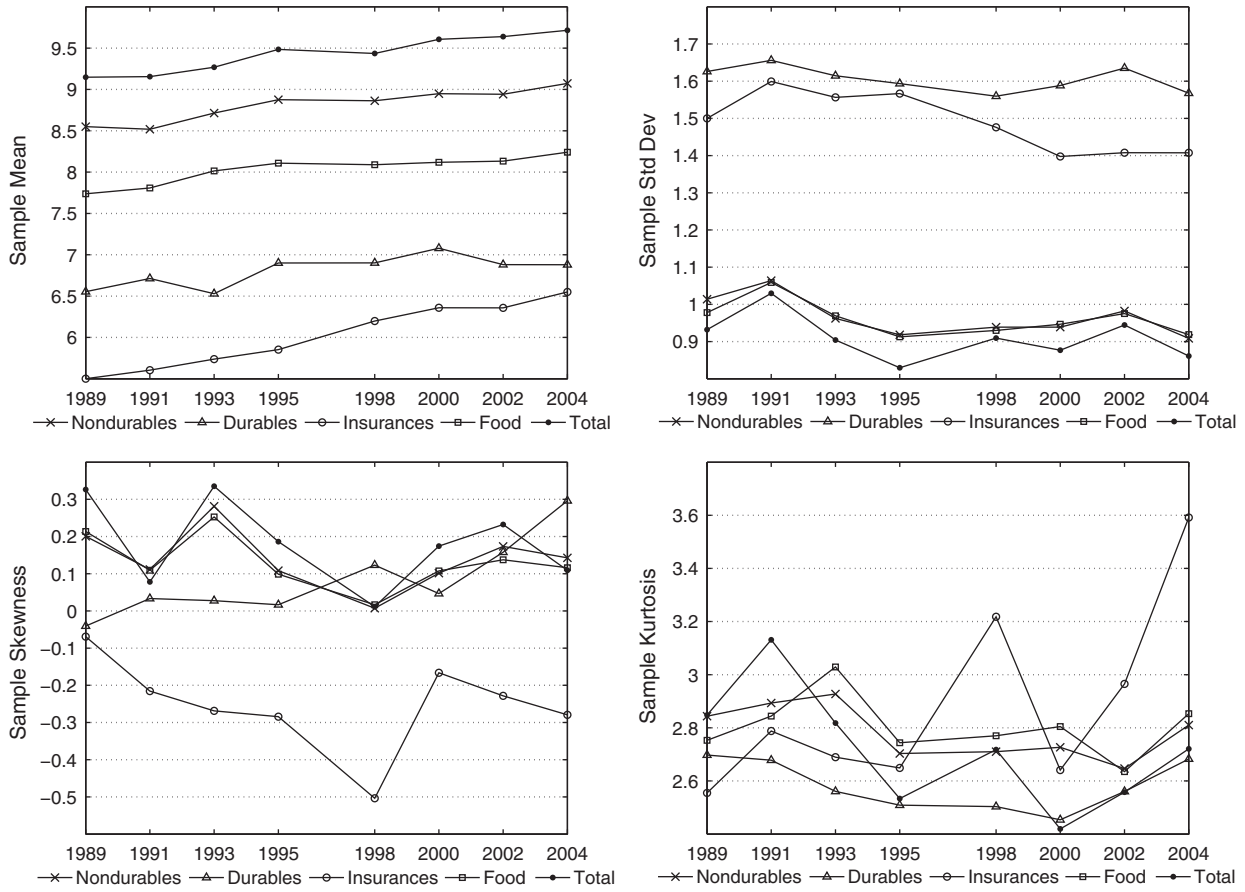

Fig. 3. Sample moments of logged household consumption expenditure distributions and their evolution over time.

of household budget shares with the sum of nondurables, durables and insurance expenditure (i.e., by replacing the variable total expenditure provided by the Bank of Italy with the sum of the commodity categories of expenditure employed in this study).

Fig. 6 plots the time evolution of the first four sample moments of household budget share distributions. In general, moments are relatively stable over time. The exception is represented again by insurance premia, which display highly increasing moments from 1989 on. In particular, skewness and kurtosis exhibit big jumps in 1995, and then move to higher levels. Instead, standard deviation steadily increases from 1989 on. Notice also that skewness signs do not change over time: they are always negative for nondurable goods and always positive for all other categories (see Table 3).

The main message coming from the foregoing visual analysis is that household consumption and budget share distributions did not dramatically change their structural properties over time. However, all previous kernel plots were not complemented by their confidence intervals. This prevents one to correctly compare kernel density estimates and to assess whether kernels for successive years are really different. To better explore this point, we plot kernel-estimates 95% confidence bands for both consumption levels and budget shares in the first (1989) and final (2002) wave.[6] If confidence intervals do not overlap, we can conclude that there is some statistical evidence for non-constant kernel densities over time, at least as far as the change between 1989 (first wave) and 2002 (last wave) is concerned.

Fig. 7 shows some examples of confidence bands plots for both household consumption log-levels and budget shares. Due to space constraints, we only plot durables and non-durables, but our main results hold also for the other two categories. This exercise shows that indeed confidence intervals overlap only for some intervals of the domain, i.e., the density estimates are not statistically constant over time over the entire ranges. Nevertheless, the density estimates seem to preserve over time some properties related to the shape of the distribution, properties that are idiosyncratic to the type of commodity considered.

[6] Confidence intervals have been estimated with the option *ci* of the Stata function *kdens*, with 100 bootstrap replications. Some parts of the lower bound of the confidence interval do not show up in the plots because they are associated to very small values in the logarithmic scale.

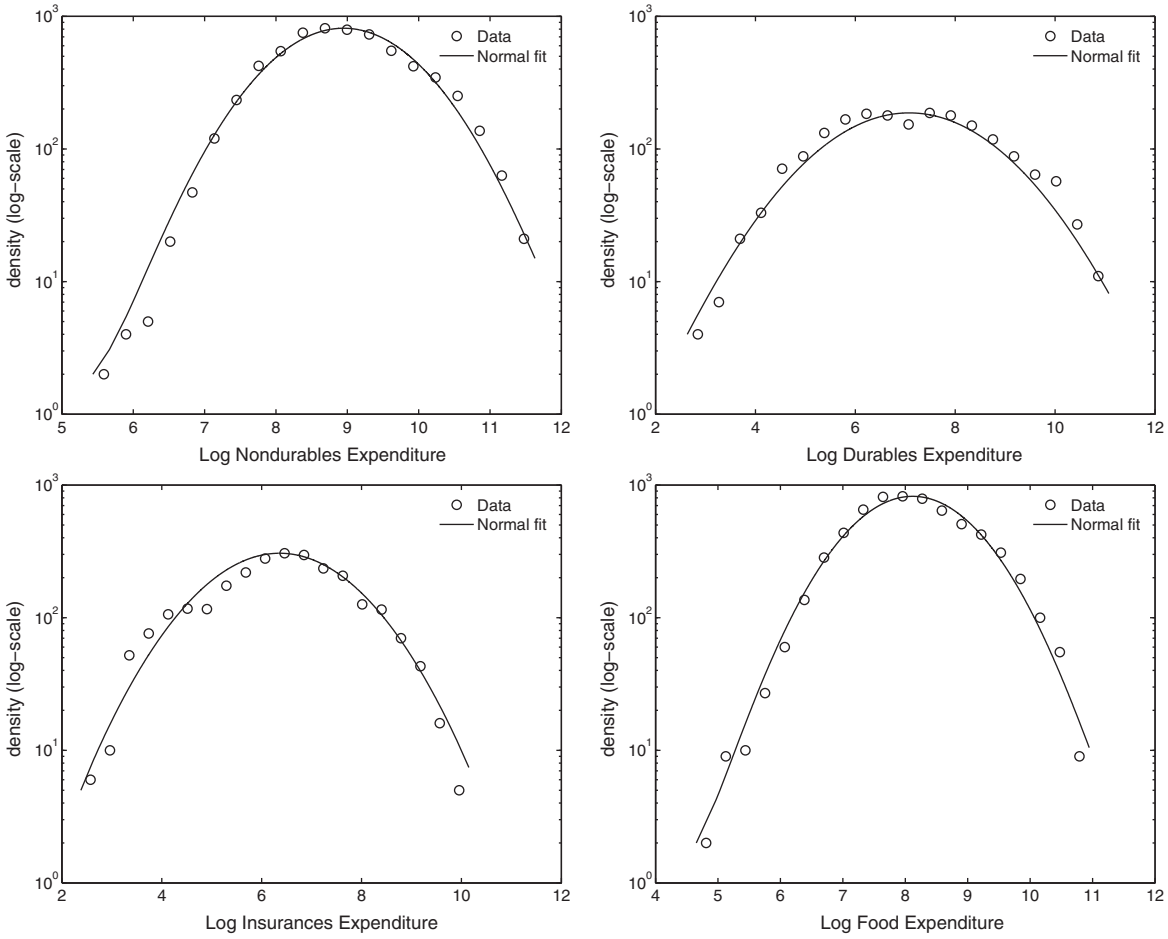

Fig. 4. An example of normal fits to logged household consumption expenditure distributions. Wave, 2000.

In other words, most of the characteristics of the empirical distributions (relative position of the mode, etc.) seem to depend more on the commodity category under analysis and not on the particular cross-sectional wave. Therefore, we can conclude that our cross-wave comparisons bring evidence in favor of the need for finding a theoretical distribution that can accommodate different budget shares, which are heterogeneous across goods much more than over time.

The relative constancy of the shape of household expenditure and budget-share distributions is a strong result also in light of the introduction of the Euro in 2001 and the increase in average nominal (and real) individual consumption levels experienced in the observed period and already noticed above. Our results that, despite a common trend due to economic growth that pushed up average consumption levels, there was a balanced turbulence within each distributions, with some households moving towards the upper tail and others moving towards the lower tail, without however changing very much the shape of the distribution. In other words, the overall impact of macroeconomic dynamics on the shape of the micro distribution was relatively weak.

We now turn to study the correlation structure of household budget share distributions. Table 4 reports the correlation matrix for 2004, together with the p-values for the null hypothesis of no correlation. Fig. 8 plots instead the time evolution of the correlations between the distributions of nondurable goods, durable goods and insurance premia. Note that all correlations are fairly stable over time and exhibit signs consistent with the economic intuition. Indeed, nondurables are negatively correlated with durables – the average correlation being −0.54% – which in turn are negatively correlated with food (here the average correlation coefficient is −0.3%). Negative correlations indicate that when households increase their relative expenditure for durable goods, they tend to reduce their relative expenditure for nondurable goods, including food. Notice also that the correlation between insurance and all other categories is statistically non significant.

A remark is in order. Correlation analysis is known to be subject to many difficulties in the case of compositional data (Aitchison, 1986, chapter 3.3), mainly due to the presence of the linear constraint imposing that the sum of budget shares must be one. Notice, however, that in our case this constraint does not apply as the denominator of

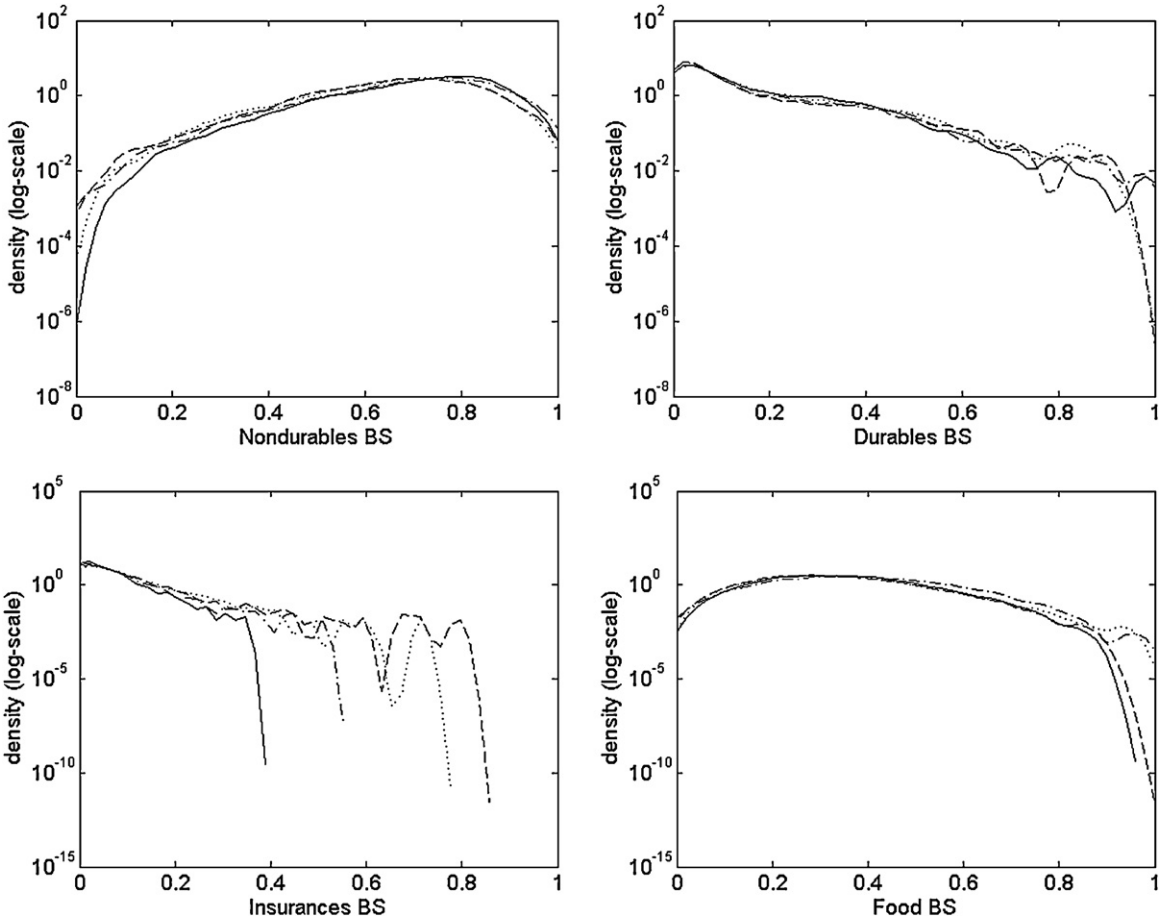

Fig. 5. Kernel-density estimates of household budget share distributions: evolution over time. Solid line, year 1989; dashed-dotted line, year 1993; dotted line, year 1998; dashed line, year 2002.

B's is not obtained as the sum of household consumption expenditure for N, D, and I (see above).

We have also performed a robust-moment analysis, using median and mean absolute deviation as robust estimators for location and scale parameters, in addition to the robust-skewness estimator already introduced above. Results confirm, overall, our previous findings: robust moments for (logged) household consumption expenditure and budget share distributions are stable over time, with the same exceptions found before.

4. A multivariate parametric approach

The most direct strategy to determine a parametric model that is able to satisfactorily fit budget-share distributions would be to employ the representation in (2) and (3), and follow a multivariate approach. Since $\underline{B}^{h,t}$, $h = 1, \ldots, H$ lies in the simplex for each wave, it is straightforward to apply standard techniques of compositional data analysis. Unfortunately, this line of attack faces in our case a number of difficulties. First, as noted in Mateu-Figueras et al. (2007a), "there is a dearth of suitable models with which to adequately model compositional data sets" (p. 217), and some of them (i.e., the Dirichlet class) are straightforwardly rejected by the data. Indeed, the Dirichlet class (Connor, 1969) requires one to assume that original household consumption expenditure data are independent and all the off-diagonal elements of the correlation matrix are negative (see Aitchison, 1986, Chapter 3). Second, and most important here, almost all viable alternatives rely on mathematical transformations of the original budget-share vector $\underline{B}$ that, albeit very useful for statistical purposes, are not able to preserve the original economic meaning of budget-shares as fractions of household total expenditure devoted to a particular class of commodities. Indeed, the basic idea of compositional analysis is to transform compositional data (i.e., share vectors lying on the simplex) in such a way to obtain vectors of quantities defined on real spaces.

This in the case also of the most simple of such transformations, namely the additive logistic ratio (ALR), which in our case would map the vector $\underline{B} = (B_{\mathrm{N}}, B_{\mathrm{D}}, B_{\mathrm{I}}, B_{\mathrm{O}})$ lying in the simplex, into the vector:

$$\mathrm{ALR}(\underline{B}) = \log(\underline{B}_{-V}/B_V) \in \mathbb{R}^3, \tag{4}$$

where V is one of the 4 original components $\{\mathrm{N}, \mathrm{D}, \mathrm{I}, \mathrm{O}\}$ and $\underline{B}_{-V}$ denotes $\underline{B}$ without the Vth component.

Table 3
Moments of household budget share distributions vs. waves. Avg, average values over the whole period; N, nondurables; D, durables; I, insurances; F, food. The figures labeled as N + D+ I only refer to households with non-zero expenditure for each commodity category.

	Stats	Waves								Avg
		1989	1991	1993	1995	1998	2000	2002	2004	
N	*N* obs.	7409	7208	6223	6258	5588	6277	6361	6281	6451
	Mean	0.717	0.702	0.703	0.699	0.667	0.675	0.668	0.666	0.687
	Std. dev.	0.141	0.154	0.151	0.142	0.149	0.143	0.147	0.146	0.147
	Skewness	−0.873	−0.736	−0.797	−0.686	−0.698	−0.726	−0.652	−0.610	−0.722
	Kurtosis	3.674	3.199	3.583	3.425	3.317	3.508	3.500	3.339	3.443
D	*N* obs.	2534	2352	2082	1856	2091	1920	1833	1961	2079
	Mean	0.130	0.148	0.118	0.127	0.137	0.136	0.118	0.105	0.127
	Std. dev.	0.138	0.150	0.143	0.144	0.151	0.149	0.144	0.137	0.144
	Skewness	1.640	1.591	1.967	1.752	1.755	1.648	1.994	2.233	1.822
	Kurtosis	5.767	5.663	7.182	6.139	6.146	5.610	7.072	8.207	6.473
I	*N* obs.	1780	1928	2257	2961	2652	2575	2175	2164	2312
	Mean	0.039	0.043	0.048	0.049	0.062	0.062	0.060	0.066	0.054
	Std. dev.	0.040	0.042	0.051	0.057	0.063	0.066	0.067	0.079	0.058
	Skewness	2.116	1.799	2.609	3.954	2.544	3.270	3.700	4.281	3.034
	Kurtosis	9.798	7.403	14.980	39.336	14.909	23.049	27.494	33.127	21.262
F	*N* obs.	7409	7228	6235	6261	5596	6281	6366	6281	6457
	Mean	0.335	0.364	0.369	0.343	0.323	0.310	0.314	0.305	0.333
	Std. dev.	0.122	0.139	0.138	0.127	0.124	0.117	0.124	0.117	0.126
	Skewness	0.405	0.402	0.285	0.389	0.557	0.457	0.530	0.586	0.452
	Kurtosis	3.051	2.938	2.863	3.018	3.563	3.373	3.218	3.423	3.181
N + D + I	*N* obs.	896	904	1099	1225	1310	1162	930	1016	1069
	Mean	0.805	0.790	0.794	0.809	0.815	0.817	0.803	0.798	0.804
	Std. dev.	0.153	0.159	0.175	0.150	0.151	0.158	0.142	0.172	0.157
	Skewness	−0.423	−0.129	−0.083	0.125	0.220	0.258	0.129	0.741	0.105
	Kurtosis	4.790	5.211	5.079	5.173	4.884	6.287	4.888	6.382	5.337

Note that the choice of V among {N, D, I, O} does not affect, from a statistical point of view, the goodness-of-fit of the parametric model chosen for representing the data (see Aitchison, 1986, Chapter 7), but introduces a stringent trade-off as far as economic interpretation is concerned. On the one hand, a first obvious choice would be to set $V = O$ and study the log-ratio vector $\log(B_N/B_O, B_D/B_O, B_I/B_O) = \log(C_N/C_O, C_D/C_O, C_I/C_O)$. This would keep the focus on the three most important commodity categories, but the interpretability of the results in terms of the "O" category defined ex-post would be strongly reduced. On the other hand, letting $V \in \{N, D, I\}$ would improve the economic interpretation, as log-ratios would represent excess of expenditure for one commodity category in terms of another meaningful category on the $(-\infty, +\infty)$ support (with 0 standing for equal expenditure). However, setting $V \in \{N, D, I\}$ would imply to lose one important dimension of the analysis – i.e., the study of the shape of one important expenditure category for which we have reliable data – which would be instead preserved with the choice $V = O$. For this reason, we have chosen here to present all results with $V = O$. In both cases, the original setup in terms of well-defined economic budget-share variables is lost.

Despite this lack of interpretability of results, we have attempted to fit our data with the two most-widely employed multivariate distributions that employ ALR transformations (Aitchison and Egozcue, 2005), namely the additive-logistic multivariate normal (ALN) distribution (Aitchison, 1986, Chapter 6) and the additive-logistic multivariate skew-normal (ALSN) distribution (Mateu-Figueras et al., 2007a; Mateu-Figueras and Pawlowsky-Glahn, 2007). Formally, $\underline{B}$ is ALN-distributed if $\mathrm{ALR}(\underline{B})$ is multivariate normally distributed. Notice that this necessarily applies if (C_N, C_D, C_I, C_O) is multivariate log-normally distributed, which does not seem to be the case for our data according to a multivariate normality Energy-test (see Section 3.1). Similarly, $\underline{B}$ is ALSN-distributed if $\mathrm{ALR}(\underline{B})$ is a multivariate skew normal (see Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999, for a formal definition). In other words, the multivariate skew-normal distribution is an extension of the multivariate normal allowing for a moderate (positive or negative) skewness.

Table 5 summarizes our results for wave 2004 (but findings are very similar in all the waves). As the first three lines of the table show, normality of $\underline{B}$ is rejected both jointly – according to the Energy test – and separately for each marginal according to both Jarque–Bera (Bera and Jarque, 1980, 1981) and Lilliefors (Lilliefors, 1967). This implies that ALN does not seem to provide a good multivariate description of our ALN-transformed data. Notice, however,

Table 4
Correlations among household budget share distributions and *p*-values (in brackets) for the null hypothesis of no correlation. Wave 2004. N, nondurables; D, durables; I, insurances; F, food.

	N	D	I
D	−0.55	–	–
	(0.00)	–	–
I	0.02	0.05	–
	(0.44)	(0.08)	–
F	0.48	−0.31	0.04
	(0.00)	(0.00)	(0.17)

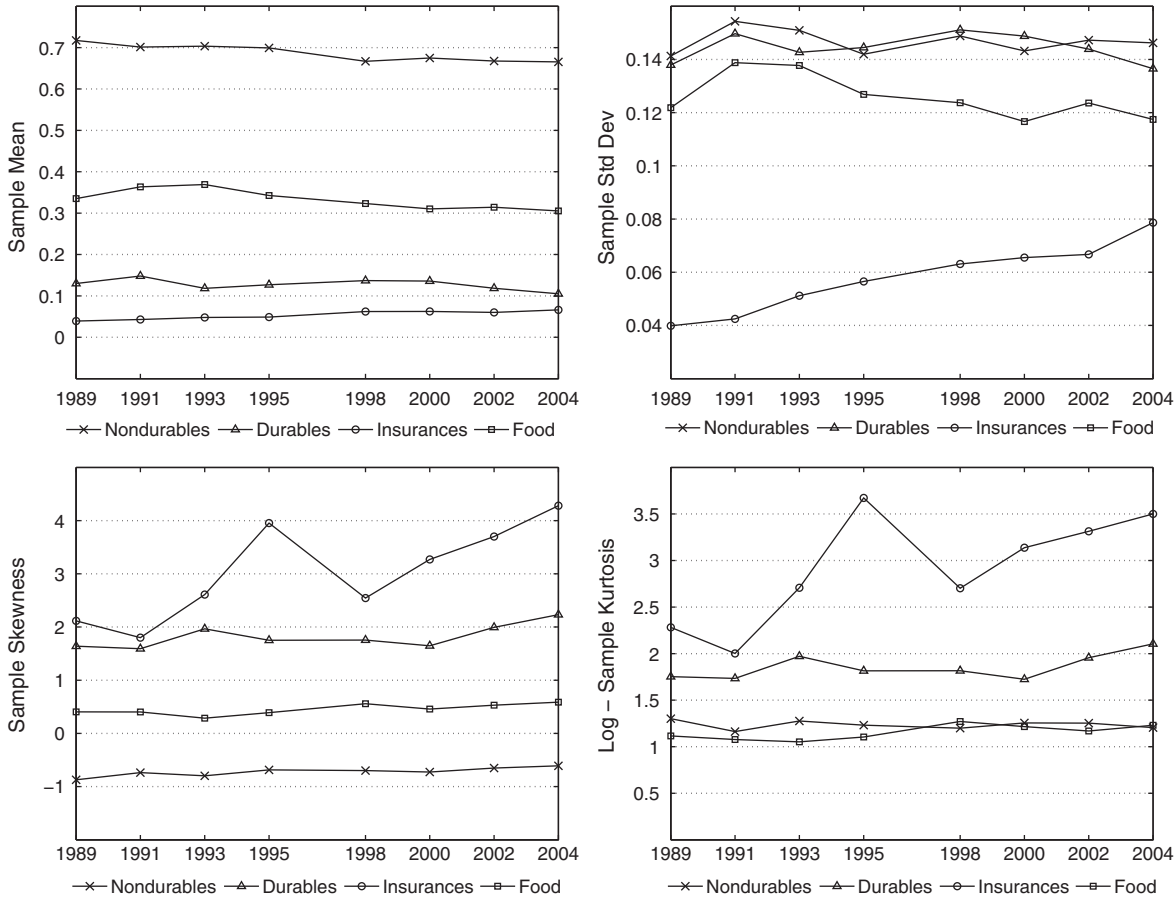

Fig. 6. Sample moments of household budget share distributions and their evolution over time.

that in addition to the third standardized moment (i.e., 3rd moment about the mean divided by the standard deviation σ), also the γ_1 coefficient (defined as $\gamma_1 = 0.5(4-\pi)m_3/\sigma^3$, where m_3 is the third moment), the D'Agostino skewness test (D'Agostino, 1970), and the robust-skewness estimator of Groeneveld and Meeden (1984) all suggest some (moderate) positive skewness. As mentioned, a skew-normal distribution can only accommodate moderate skewness, i.e., values of γ_1 in the interval $[-0.995, 0.995]$ (Azzalini and Dalla Valle, 1996). It is thus natural to check whether the

Table 5
Multivariate analysis for the log-ratio vector $\log(B_N/B_O, B_D/B_O, B_I/B_O)$, where N, nondurables; D, durables; I, insurances; F, food; Wave, 2004. Sample Skewness: 3rd moment about the mean divided by the standard deviation. γ_1 index defined as $0.5(4-\pi)m_3/\sigma^3$, where m_3 is the 3rd moment. The d-test (Mateu-Figueras et al., 2007b) for multivariate skew normality is distributed as a χ^2_3.

Statistic	Standardized log-ratios		
	$\log(B_N/B_O)$	$\log(B_D/B_O)$	$\log(B_I/B_O)$
Multivariate normality			
Energy test		4.9518 (0.0000)	
Univariate normality			
Lilliefors	0.0529 (0.0010)	0.0725 (0.0010)	0.0305 (0.0514)
Jarque-Bera	627.2247 (0.0000)	56.0317 (0.0000)	65.5013 (0.0000)
Skewness			
Sample skewness	0.8702	0.5843	0.0408
γ_1 index	0.3728	0.2504	0.0175
D'Agostino test	6.0741 (0.0000)	4.3550 (0.0000)	0.3267 (0.0741)
Robust skewness	0.0574 (0.1049)	0.2138 (0.0000)	−0.0689 (0.0697)
Multivariate skew-normality			
d-Test		35.2373 (0.0000)	
Univariate skew-normality			
Quadratic AD	16.2662 (0.0010)	10.1291 (0.0175)	15.8801 (0.0012)

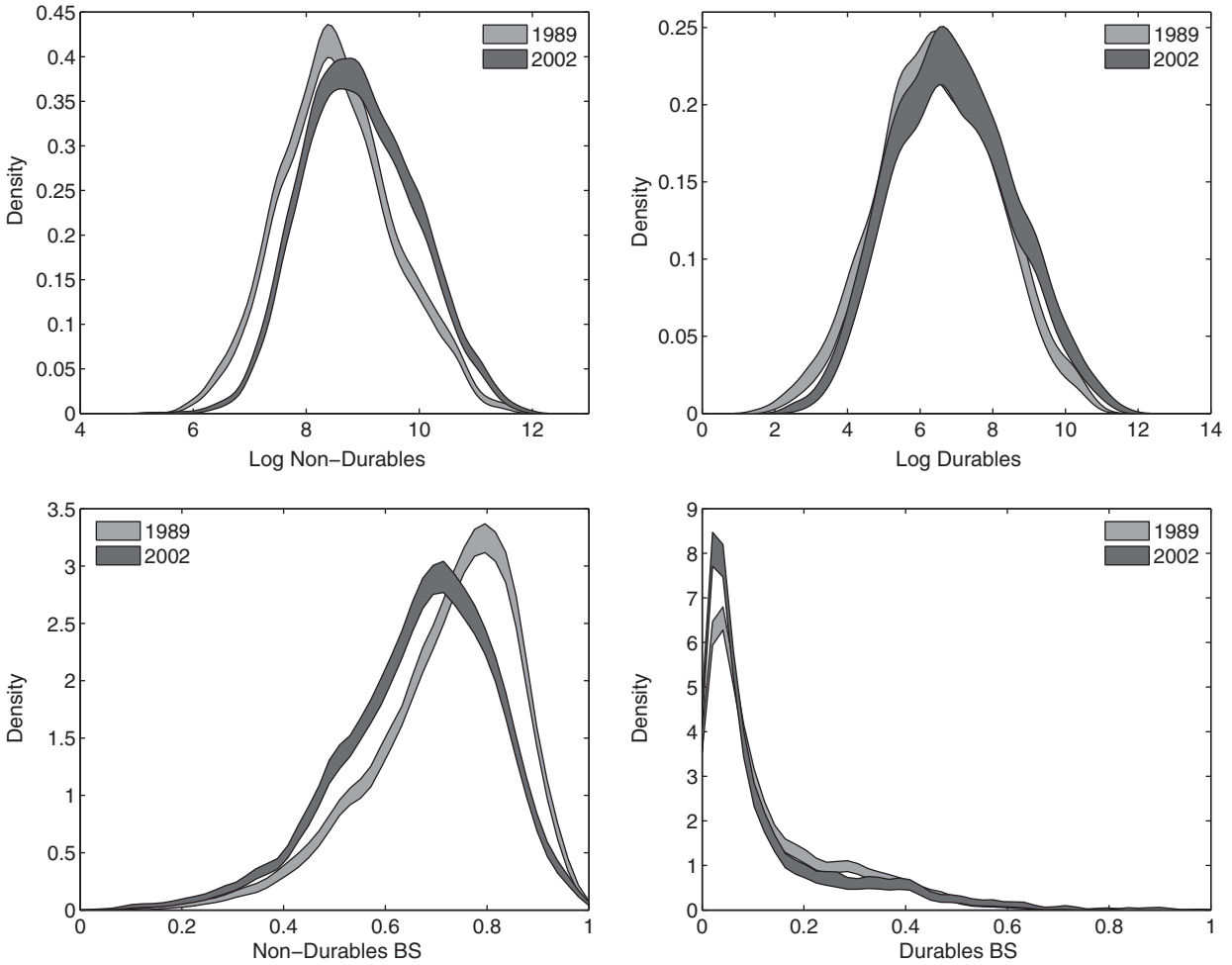

Fig. 7. 95% Confidence bands for kernel estimates. Comparison between first (1989) and final (2002) wave. Top two panels: household expenditure distributions. Bottom two panels: household budget-share distributions.

joint $\underline{B}$ distribution can be satisfactorily proxied by a multivariate ALSN distribution. This amounts to test if, jointly, ALR($\underline{B}$) is multivariate skew-normal. We employ here the d test discussed in Mateu-Figueras et al. (2007b), which under the null of multivariate skew-normality should be distributed here as a χ^2_3. As Table 5 shows, the null is rejected, as also happens for the null hypotheses of skew-normality of the three marginal log-ratio distributions,

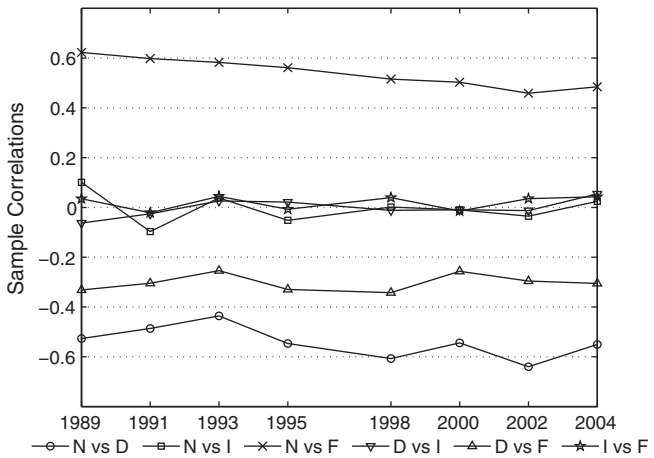

Fig. 8. Correlations between household budget share distributions and their evolution over time. N, nondurables; D, durables; I, insurances; F, food.

according to standard quadratic Anderson–Darling tests (see D'Agostino and Stephens, 1986). Of course, rejection by Anderson–Darling EDF-based tests may simply be due to a size effect, as our samples are always very large (see, e.g., Bentler and Bonett, 1980). To control for such a problem, we have run GoF tests on binned data and on randomly drawn smaller-sized sub-samples without noticing any major differences in our results. We shall go back to this point in Section 6.

5. Towards a univariate parametric model for budget share distributions

The foregoing analysis suggests that – from a purely statistical perspective – our data cannot be easily proxied by the mostly employed parametric multivariate distributions defined on the simplex. More importantly, such representations often require mathematical transformations of the data that would strongly reduce the possibility of economically interpreting the results, let aside the difficulty of treating sub-categories as food. In what follows, we will therefore turn to a univariate analysis aiming at determining a parsimonious, parametric, model able to satisfactorily fit household budget share (marginal) distributions, without using mathematical transformations of the original variables that may undermine the economic interpretations of the results.

We look for a family of univariate densities, defined on the unit interval, holding at least the following three desirable features. First, the family of densities fitting household budget shares should be consistent with the statistical properties of the underlying household consumption expenditure univariate distributions employed to compute budget shares. Second, it should be flexible enough to accommodate – for each wave – the observed across-commodity heterogeneity in the shape of household budget share distributions. Third, the parameters of the density should embody some economic meaning and allow one to taxonomize commodity categories according to their (high and low) level.

Let $C_1, \ldots, C_K$ be the expenditure levels of a given household in a representative time period, where K is the number of commodity categories considered. The household budget share of commodity category i is defined as

$$B_i = \frac{C_i}{C} = \frac{1}{1 + (\sum_{j \neq i} C_j)/C_i} = \frac{1}{1 + \sum_{j \neq i} Z_j(i)} = \frac{1}{1 + S_i} \quad (5)$$

where S_i is the sum of the $K-1$ random variables $Z_j(i)$, each being equal to the ratio between C_j and C_i, with $j = 1, \ldots, i-1, i+1, \ldots, K$. Obviously, $B_i \in (0, 1)$ as required. From Eq. (5), it follows that the cumulative distribution function (cdf) of B_i reads:

$$F_{B_i}(x) = \text{Prob}\{B_i < x\} = \text{Prob}\left\{1 + S_i > \frac{1}{x}\right\} = 1 - F_{S_i}\left(\frac{1}{x} - 1\right), \quad (6)$$

where $x \in (0, 1)$ and F_{S_i} is the cdf of S_i. Therefore, the probability density function (pdf) of B_i is given by:

$$f_{B_i}(x)\mathrm{d}x = \frac{1}{x^2} f_{S_i}\left(\frac{1}{x} - 1\right) \mathrm{d}x, \quad (7)$$

where f_{S_i} is the pdf of S_i. This means that characterizing the distribution of B_i requires studying the distribution of $S_i = \sum_{j \neq i} C_j / C_i = \sum_{j \neq i} Z_j(i)$. Given the empirical evidence above, there are good reasons to assume that expenditure levels C_i are all log-normally distributed, at least as a first approximation. This implies that the ratios $Z_j(i)$ are also log-normally distributed, as:

$$\begin{aligned}\text{Prob}\{Z_j(i) < z\} &= \text{Prob}\{\log(C_j) - \log(C_i) < \log(z)\} \\ &= \text{Prob}\{D_j(i) < \log(z)\}. \end{aligned} \quad (8)$$

Since $\log(C_j)$ and $\log(C_i)$ are normally distributed (and possibly correlated), their difference $D_j(i)$ will also be normal. Hence $\exp(D_j(i))$ will be log-normally distributed.

As a result, the shape of household budget share distribution B_i fully depends on the shape of the sum of the $K-1$ log-normal variates $Z_j(i)$s. Notice that in general $Z_j(i)$ will not be uncorrelated. Indeed, the C_is may be correlated because of household preferences. This seems to be the case from our empirical evidence, as we have already noticed statistically significant correlations between household consumption expenditure distributions (see Table 2). The significant correlation between household consumption expenditure distributions thus implies that $Z_j(i)$ – as well as household budget share distributions – will not be independent.

According to the literature, there does not exist a closed form for the pdf of a sum of log-normal (correlated or uncorrelated) random variables and only approximations are available, see Beaulieu et al. (1995) for the case of independent summands and Mehta et al. (2006) for the case of correlated summands. The baseline result is that the distribution of S_i can be well approximated by a log-normal distribution, whose parameters depend in a non-trivial way on the parameters of the log-normals to be summed up and their covariance matrix. Many methods are available to find approximations to the parameters of the resulting log-normal distribution, see, e.g., Fenton (1960), Schwartz and Yeh (1982) and Safak and Safak (1994). We are not interested here in this issue because we can directly estimate the parameters of the resulting distribution for B_i via maximum likelihood.

The log-normal proxy to the sum of log-normals is not, however, the only approximation available. Indeed Milevski and Posner (1998) show that when $K \to \infty$ then S_i converges in distribution to an an inverse-Gamma ($\text{Inv}\Gamma$) density, which performs well in approximating the sum also for very small K. More formally, the $\text{Inv}\Gamma$ random variable is defined as the inverse of a Γ random variable, i.e., if $X \sim \Gamma(\theta, p^{-1})$ then $X^{-1} \sim \text{Inv}\Gamma(\theta, p)$. Therefore there may be some gains in considering an $\text{Inv}\Gamma$ proxy to S_i instead of a log-normal one. Of course, the extent to which either approximation is to be preferred is an empirical issue. For this reason, we shall consider both proxies in our empirical application below.

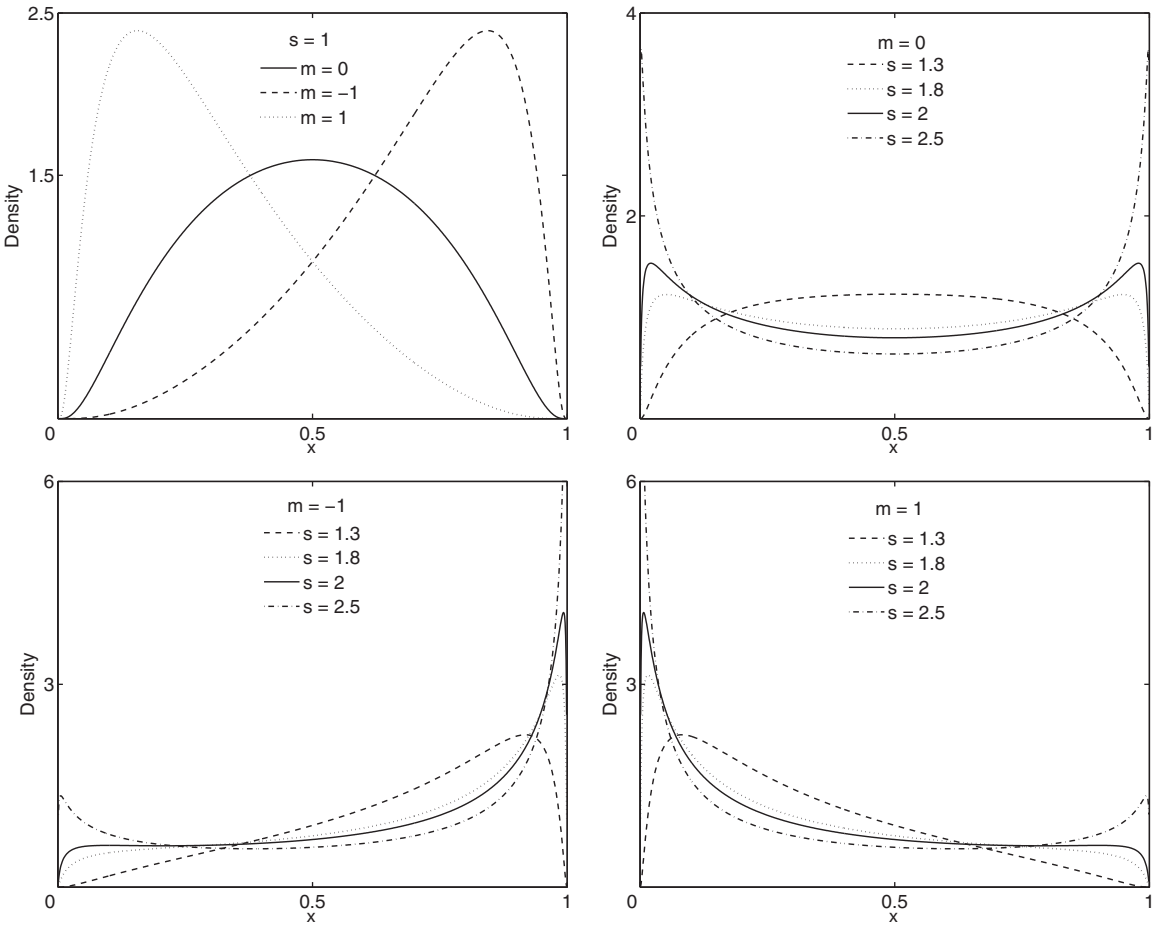

Fig. 9. The LN-B approximation to household budget share distributions. Different shapes of f_{B_i} as parameters m and s change.

In the case S_i has a log-normal pdf with parameters (m, s), then:

$$f_{S_i}(x; m, s) = \frac{1}{xs\sqrt{2\pi}} \exp\left[-\frac{(\log(x) - m)^2}{2s^2}\right]. \tag{9}$$

Using (7), we get:

$$f_{B_i}(x; m, s) = \frac{1}{x(1-x)s\sqrt{2\pi}} \exp\left[-\frac{(\log(1-x) - \log(x) - \log(m))^2}{2s^2}\right] \tag{10}$$

In what follows we shall refer to density (10) as the LN-B density. Note that the LN-B is already a pdf given that its integral over [0, 1] is one. In Fig. 9 we show a variety of shapes derived from (10) for selected values of the parameters m and s. If $m > 0$ ($m < 0$) the distribution is right-skewed (left-skewed), if $m = 0$ it is symmetric. If $0 < s \leq 1.5$ the distribution is bell-shaped, if $1.5 < s \leq 2.5$ it is bimodal, while if $s > 2.5$ it is U-shaped. This seems to confirm that despite its parsimony, the density (10) is sufficiently flexible to accommodate different shapes for household budget share distributions.

On the other hand, if we assume an InvΓ approximation for the distribution of a sum of log-normals, then the distribution of S_i depends on two parameters (θ, p) and its pdf reads:

$$f_{S_i}(x; \theta, p) = \frac{\theta^p}{\Gamma(p)} x^{-p-1} \exp\left[-\frac{\theta}{x}\right] \tag{11}$$

Once again, using (7) we obtain the pdf of B_i (henceforth, InvΓ-B), which reads:

$$f_{B_i}(x; \theta, p) = \frac{\theta^p}{x^2\Gamma(p)} \left(\frac{1}{x} - 1\right)^{-p-1} \exp\left[-\theta\frac{x}{1-x}\right]. \tag{12}$$

Fig. 10 shows the shape of the density (12) for selected values of θ and p. We immediately see that (12) is always an asymmetric distribution, as $f_{B_i}(1; \theta, p) = 0$ for any values of the parameters, while if $p > 1$ $f_{B_i}(0; \theta, p) = 0$ but if $p \leq 1$ $f_{B_i}(0; \theta, p) > 0$. The interpretation of the two parameters is less straightforward than in the previous case. Notice that for small values of p the function is monotonically decreasing, while as p increases a rightward-shifting maximum emerges. When $p < \theta$ ($p > \theta$) the maximum is attained for $x < 0.5$ ($x > 0.5$), while if $p = \theta$ the maximum is around $x = 0.5$: this is the most symmetric case we can model with this distribution. Even if the proxy (12) seems to be less flexible

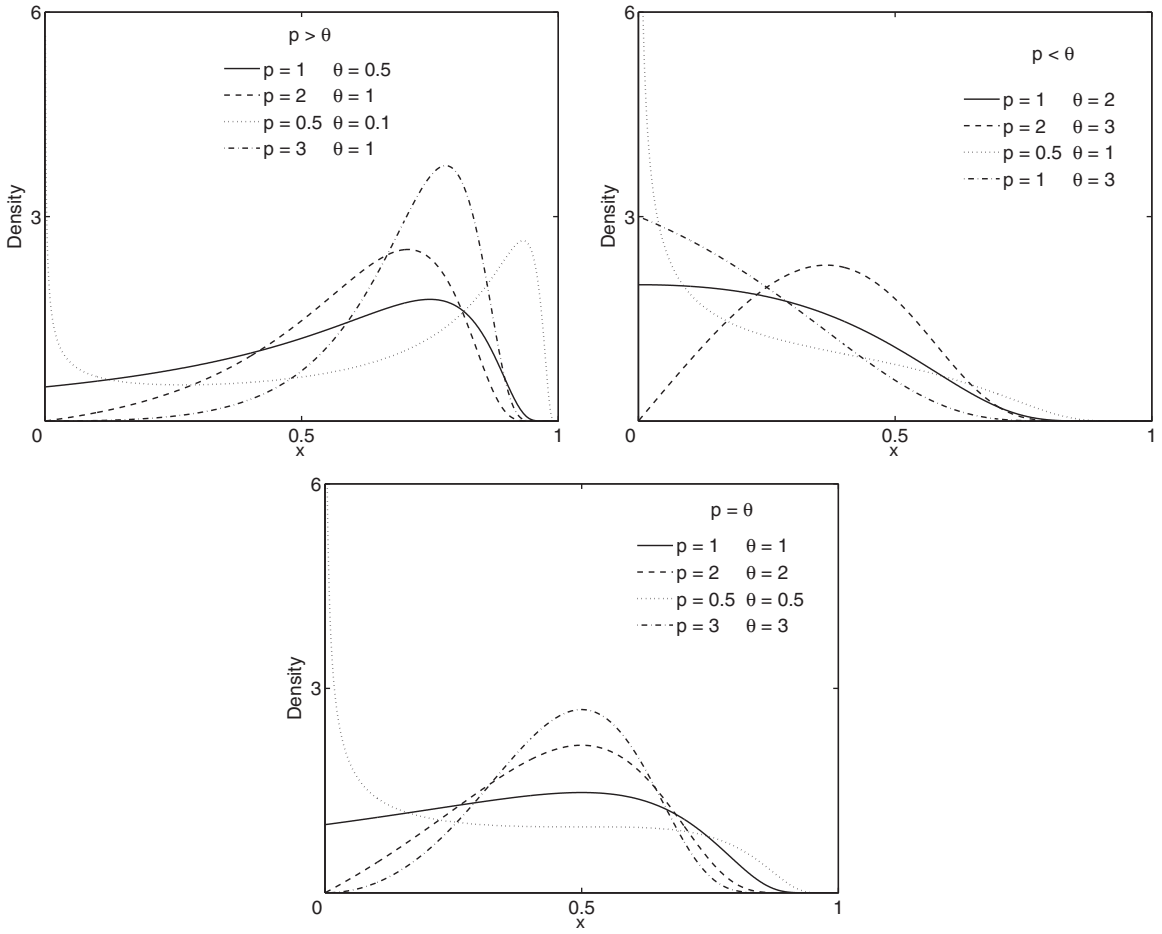

Fig. 10. The InvΓ-B approximation to household budget share distributions. Different shapes of f_{B_i} as parameters p and θ change.

than (10), we shall retain it in our fitting exercises for the sake of comparison.

6. Measuring the goodness of fit

In the previous section, we have derived two alternative, parsimonious, approximations of univariate budget-share distributions, which appear – at least in principle – flexible enough to accommodate the observed shape heterogeneity and are consistent with the empirically detected log-normality of household consumption expenditure distributions.

To check how well the foregoing approximations fit the data, we firstly estimate the parameters of (10) and (12) via maximum likelihood. Results are reported in Table 6, together with asymptotic standard deviations for the parameters of LN-B and InvΓ-B. We shall comment parameter estimates in Section 7, where they will be employed to classify the commodity categories under study. In the rest of this section, we focus instead on goodness-of-fit considerations.

To evaluate the performance of the two proposed proxies in fitting household budget share distributions as compared to alternative distributions, we choose as a benchmark the univariate Beta density (Evans et al., 2000), whose pdf reads:

$$b(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{BE(\alpha,\beta)}, \tag{13}$$

where $x\in[0,1]$ and BE is the Beta function. Notice that the Beta also depends on only two parameters and typically is flexible enough to accommodate many alternative shapes. However, it lacks any consistency requirements with respect to the underlying shape of household consumption expenditure distributions, because in general it cannot be derived as the density of household budget shares stemming from log-normally distributed expenditure levels.

We firstly employ a goodness-of-fit tests based on empirical distribution function statistics. More specifically, we run three widely used tests: Kuiper (KUI; Kuiper, 1962), Cramér-von Mises (CvM; Pearson and Stephens, 1962) and quadratic Anderson–Darling (AD2; Anderson and Darling, 1954). We are interested in understanding whether the theoretical distributions we suggest are able to proxy the empirical data better than plausible alternatives, because of their consistency with the findings above on household consumption expenditure distributions. Notice that,

Table 6
Estimated parameters and asymptotic standard deviations (in parentheses) of LN-B and InvΓ-B vs. waves. N, nondurables; D, durables; I, insurances; F, food.

LN-B		Waves							
		1989	1991	1993	1995	1998	2000	2002	2004
N	m	−1.04 (0.009)	−0.98 (0.010)	−0.98 (0.010)	−0.95 (0.010)	−0.77 (0.010)	−0.81 (0.009)	−0.78 (0.009)	−0.77 (0.009)
	s	0.76 (0.006)	0.83 (0.007)	0.83 (0.007)	0.77 (0.007)	0.74 (0.007)	0.72 (0.006)	0.75 (0.007)	0.74 (0.007)
D	m	2.47 (0.026)	2.28 (0.027)	2.69 (0.031)	2.57 (0.033)	2.43 (0.030)	2.45 (0.031)	2.64 (0.032)	2.80 (0.030)
	s	1.33 (0.019)	1.33 (0.019)	1.42 (0.022)	1.40 (0.023)	1.37 (0.021)	1.37 (0.022)	1.35 (0.022)	1.34 (0.021)
I	m	3.74 (0.028)	3.66 (0.028)	3.61 (0.028)	3.61 (0.024)	3.25 (0.024)	3.22 (0.023)	3.26 (0.025)	3.17 (0.026)
	s	1.18 (0.020)	1.24 (0.020)	1.31 (0.020)	1.31 (0.017)	1.25 (0.017)	1.18 (0.016)	1.17 (0.018)	1.20 (0.018)
F	m	0.74 (0.007)	0.61 (0.008)	0.59 (0.008)	0.71 (0.008)	0.80 (0.008)	0.86 (0.008)	0.85 (0.008)	0.89 (0.008)
	s	0.59 (0.005)	0.66 (0.005)	0.66 (0.006)	0.61 (0.005)	0.62 (0.006)	0.60 (0.005)	0.62 (0.006)	0.60 (0.005)

InvΓ-B		Waves							
		1989	1991	1993	1995	1998	2000	2002	2004
N	p	1.79 (0.027)	1.45 (0.022)	1.38 (0.022)	1.52 (0.025)	1.93 (0.034)	1.96 (0.032)	1.74 (0.028)	1.81 (0.030)
	θ	0.47 (0.008)	0.37 (0.007)	0.34 (0.007)	0.41 (0.008)	0.67 (0.013)	0.66 (0.012)	0.58 (0.011)	0.62 (0.012)
D	p	0.66 (0.016)	0.68 (0.017)	0.57 (0.015)	0.63 (0.017)	0.64 (0.017)	0.64 (0.017)	0.61 (0.017)	0.58 (0.015)
	θ	3.09 (0.106)	2.70 (0.095)	2.78 (0.108)	3.08 (0.124)	2.81 (0.106)	2.83 (0.112)	3.12 (0.127)	3.26 (0.130)
I	p	0.99 (0.029)	0.95 (0.027)	0.86 (0.022)	0.78 (0.017)	0.92 (0.022)	0.93 (0.023)	0.91 (0.024)	0.71 (0.018)
	θ	23.14 (0.877)	20.14 (0.738)	15.95 (0.549)	13.37 (0.409)	12.69 (0.398)	12.50 (0.398)	12.46 (0.433)	7.30 (0.265)
F	p	3.07 (0.048)	2.38 (0.037)	2.52 (0.043)	2.86 (0.048)	2.73 (0.049)	3.02 (0.051)	2.77 (0.046)	3.00 (0.051)
	θ	5.41 (0.092)	3.51 (0.061)	3.68 (0.069)	4.82 (0.089)	5.00 (0.098)	6.02 (0.111)	5.34 (0.098)	6.09 (0.112)

as our samples are always very large, the null assumption that empirical data are drawn from any given theoretical distribution would bring the traditional EDF-based goodness-of-fit tests, which evaluate the deviation of the empirical CDF from the theoretical one at each single observation, to reject the null hypothesis at any probability level, despite only minor discrepancies (see Bentler and Bonett, 1980). Therefore, we shall evaluate in what follows the goodness of fit of our competing densities after having grouped logged observations among equally spaced bins and computed test statistics over such bins. In this way, the effect of a discrepancy between theoretical and empirical probability density will strongly affect the statistic only if the same discrepancy is recurring frequently within a particular interval (bin), while the existence of sporadic outliers will only have a minor effect on the test statistic. This procedure also alleviates the difficulties coming from the generation of pseudo-random numbers distributed as in (10) or (12). This is not a trivial task and the issue is one of the main points of our research agenda.

Table 7 reports the goodness-of-fit statistics only for the AD2 test, as in general all three tests agree on whether the benchmark is outperformed or underperformed by either specification of the proposed density. According to test statistics, the Beta fit is never the best one. With respect to the LN-B and the InvΓ-B densities, the latter seems to deliver a better fit in the case of nondurable and durable goods (and according to KUI and CvM tests, also in the case of food). However, the levels of the statistics do not take into account the different influences exerted by the sample size on the results of the binning procedure for different distributions. Moreover, a density may perform relatively better than another one even though both provide a very bad description of the empirical sample. In order to perform a more statistically sound comparison, we have therefore proxied via simulation the distributions of the test statistics when using our procedure of grouping the observations into bins, and we have computed the relevant p-values of the empirical values obtained before, i.e., the probability mass to the right of the observed statistics.

More specifically, to proxy the distribution of a test statistic for a given theoretical density and a commodity category of household budget shares of size n, we use the following procedure: (i) generate via a bootstrap-with-replacement method a random sample of observations of the same size n as the observed sample, then group the observations into L bins; (ii) on the randomly extracted sample, re-estimate the parameters of the theoretical frequency by maximum likelihood and compute the test statistic; (iii) repeat this procedure a large number of times m to get the proxy for the distribution of the test statistic. Of course the foregoing steps should be repeated for any given empirical sample, i.e., for any wave and commodity category considered, and for any of the three densities studied. In what follows, we have set $m = 1000$ and we have considered $L = 100$ bins.

The picture given by the p-values is generally consistent with that given by the statistics, however in a dozen cases the p-values for the Beta density are at least as good as those for the LN-B or the InvΓ-B densities, notwithstanding lower values of the statistics. Still, in terms of p-values, the LN-B or the InvΓ-B densities outperform the Beta density in the 78%, 59% and 78% of the cases according to KUI, CvM and AD2, respectively.

Together with standard goodness-of-fit tests, we employ also the average absolute deviation as a simple measure of goodness-of-fit. The average absolute deviation represents an alternative, additional, measure of

Table 7
Quadratic Anderson–Darling (AD2) statistics with p-values (in brackets). N, nondurables; D, durables; I, insurances; F, food. In boldface the figures that in any given wave and commodity category minimize statistics or maximize p-values.

	pdf	Waves							
		1989	1991	1993	1995	1998	2000	2002	2004
N	Beta	73.01 (0.33)	38.64 (0.33)	72.83 (0.32)	50.94 (0.34)	49.14 (**0.36**)	48.70 (**0.41**)	60.60 (0.35)	45.25 (**0.30**)
	LN-B	65.13 (0.34)	34.86 (0.32)	63.86 (0.32)	44.29 (0.35)	45.96 (0.35)	45.04 (0.40)	55.26 (**0.36**)	41.56 (**0.30**)
	InvΓ-B	**42.55**(**0.35**)	**24.07**(**0.37**)	**36.05**(**0.34**)	**26.42**(**0.40**)	**31.77**(0.36)	**31.05**(0.21)	**33.73**(0.34	**27.48**(0.28)
D	Beta	193.28 (**0.26**)	149.61 (0.30)	189.78 (0.26)	148.09 (0.41)	134.88 (0.36)	161.90 (0.28)	168.69 (0.35)	224.49 (0.25)
	LN-B	150.16 (0.25)	121.19 (0.29)	153.64 (0.29)	124.44 (**0.43**)	116.48 (0.38)	133.49 (**0.29**)	146.18 (0.36)	186.59 (**0.27**)
	InvΓ-B	**130.92**(0.20)	**112.83**(**0.39**)	**115.92**(**0.33**)	**112.10**(0.43)	**103.96**(**0.40**)	**117.75**(0.19	**117.01**(**0.44**)	**129.91**(0.18)
I	Beta	212.45 (0.31)	155.52 (0.29)	267.57 (0.27)	560.43 (0.30)	351.57 (0.25)	415.20 (0.28)	436.94 (0.25)	514.84 (0.25)
	LN-B	**139.00**(**0.32**)	**107.05**(**0.31**)	**158.93**(**0.30**)	274.53 (0.32)	**194.88**(**0.28**)	**234.79**(**0.29**)	**248.61**(**0.28**)	286.99 (0.26)
	InvΓ-B	185.45 (0.30)	144.63 (0.29)	204.32 (0.28)	**257.90**(**0.42**)	250.95 (0.24)	269.22 (0.25)	266.89 (**0.28**)	**215.16**(**0.57**)
F	Beta	52.79 (0.33)	69.40 (0.29)	56.04 (0.31)	46.09 (**0.34**)	89.92 (0.42)	108.23 (0.28)	62.48 (**0.31**)	0.31 (**0.38**)
	LN-B	**49.32**(0.32)	63.35 (**0.43**)	52.59 (0.35)	**43.68**(0.32)	80.06 (**0.53**)	**90.94**(**0.31**)	**57.00**(0.29)	0.29 (0.37)
	InvΓ-B	69.34 (**0.36**)	**51.47**(0.19)	**45.94**(**0.52**)	51.93 (0.24)	**74.33**(0.34)	126.02 (0.20)	66.38 (0.25)	**0.25**(0.35)

agreement between the empirical and the theoretical frequencies. For any given commodity category i and wave t (labels are suppressed for the sake of simplicity), the average absolute deviation is defined as:

$$AAD = \frac{1}{L}\sum_{l=1}^{L} |\phi_{B_i}(x_l) - f_{B_i}(x_l; \bullet, \bullet)|, \tag{14}$$

where L is the number of bins in which we group the empirical observations, and each class is identified by its midpoint x_m, in correspondence of which we compute the empirical frequency ϕ_{B_i} and the theoretical frequency f_{B_i}. The latter is obtained using Eqs. (10), (12) or (13), when parameters are replaced by their maximum-likelihood estimates.

According to the values obtained for the average absolute deviation (see Table 8), in 66% of the cases the Beta distribution is outperformed by either the LN-B or the InvΓ-B density. More precisely, in 34% of the cases the LN-B seems to deliver a better fit, whereas in 28% of the cases the InvΓ-B approximation fits better the data and in one case they have exactly the same performance. In the remaining 34% of the cases either the LN-B or the InvΓ-B density (or even both) show the same goodness-of-fit as the Beta, i.e., the proposed density always manages to do at least as well as the benchmark. According to p-values in Table 8, the Beta distribution fits the data better than the other two densities only in 19% of the cases. Conversely, the LN-B density provides a better fit in 50% of the cases, whereas in 19% of the cases the InvΓ-B approximation wins the competition. In five cases the proposed density in either specification and the benchmark have the same performance, while in the remaining couple of cases the LN-B or the InvΓ-B densities do equally better than the Beta. It is interesting to note that, according to these results, the InvΓ-B provides good fits for nondurable goods budget shares, while the LN-B works better for durable goods and insurance premia budget shares. Food budget shares seem to be well described by either the Beta or the LN-B. Notice also that both average absolute deviations and p-values are often very similar, thus empirically it seems that in some cases all alternatives may provide equally good fits. However, the distributions that we have proposed should be in our view preferred to the Beta because of their statistical consistency with the underlying household consumption expenditure distributions.

A graphical analysis of the goodness-of-fit for two waves (2000 and 2004) is provided in Figs. 11 and 12. The LN-B density provides better fits for the left tail of nondurable budget shares and, more generally, for insurances and durables. In these latter cases, however, none of the distributions considered is able to account for the few observations lying on the extreme right of the support. The InvΓ-B performs well only on the right tail of nondurable goods budget share distributions.

7. Towards a taxonomy of commodity categories

The foregoing analysis suggests that the LN-B and InvΓ-B univariate densities are a statistically satisfactory parametric model for Italian household budget share

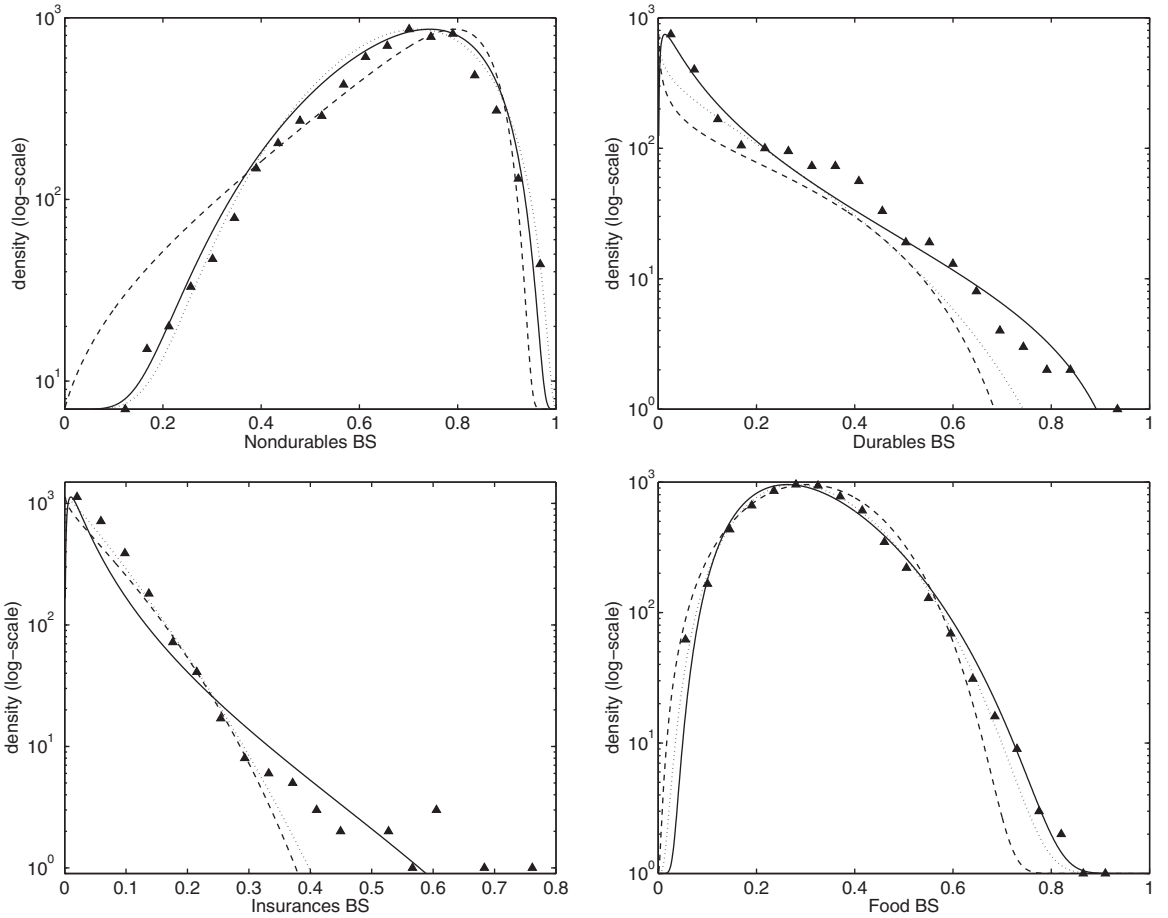

Fig. 11. Fitting alternative densities to household budget share distributions. Wave, 2000; Beta, dotted line; LN-B, solid line; InvΓ-B, dashed line.

distributions, one that is able to accommodate the existing heterogeneity in the shape of the distributions and is consistent with the statistical properties of the underlying household consumption expenditure distributions. In this section, we shall attempt to draw some economic implications stemming from estimated parameters in order to show that the family of density that we have proposed can also be employed to meaningfully classify commodity categories.

To begin with, notice that it is very hard to taxonomize our four commodity categories on the basis of the estimated parameters of household consumption expenditure distribution. Indeed, the sample moments reported in Table 1 are similar for all commodity categories. However, inspection of Table 6 reveals that estimated parameters for LN-B and InvΓ-B – as happened also for sample moments – feature a much higher heterogeneity. This difference between household budget share distributions and consumption expenditure distributions is not surprising, as household budget share distributions contain more information than household consumption expenditure distributions, namely the information about household-budget allocation behavior, which is itself the factor that can allow one to classify the commodity categories.

Therefore, it is tempting to employ the information coming from estimated parameters of both LN-B and InvΓ-B densities in order to build a taxonomy of the four commodity categories. More precisely, we shall employ the study of the shape of the LN-B and InvΓ-B densities performed in Section 5 to classify our commodity categories with respect to the high/low values of their estimated parameters (m, s) and (p, θ). Since these estimates are relatively stable across time (see again Table 6), we shall use averages of estimates across all the waves. As far as the LN-B is concerned, we shall discriminate between commodity categories exhibiting (average) estimates for $m \lessgtr 0$ and $s \lessgtr 1$, whereas for the InvΓ-B density we will differentiate between commodity categories with (average) estimates for $p \lessgtr 1$ and $\theta \lessgtr 1$. The two resulting taxonomies are shown in Table 9. Note that durable goods and insurance premia have similar characteristics, i.e., they have low dispersion and are right-skewed, while the budget share distributions of nondurable goods are more disperse and left-skewed. Food budget share distributions are similar to the latter in that are quite disperse, but are right-skewed.

Notice that, although the parameters of both the LN-B and the InvΓ-B densities cannot be easily traced back to the moments of the associated random variables, a clear-cut relation seems to exist between the taxonomies in Table 9

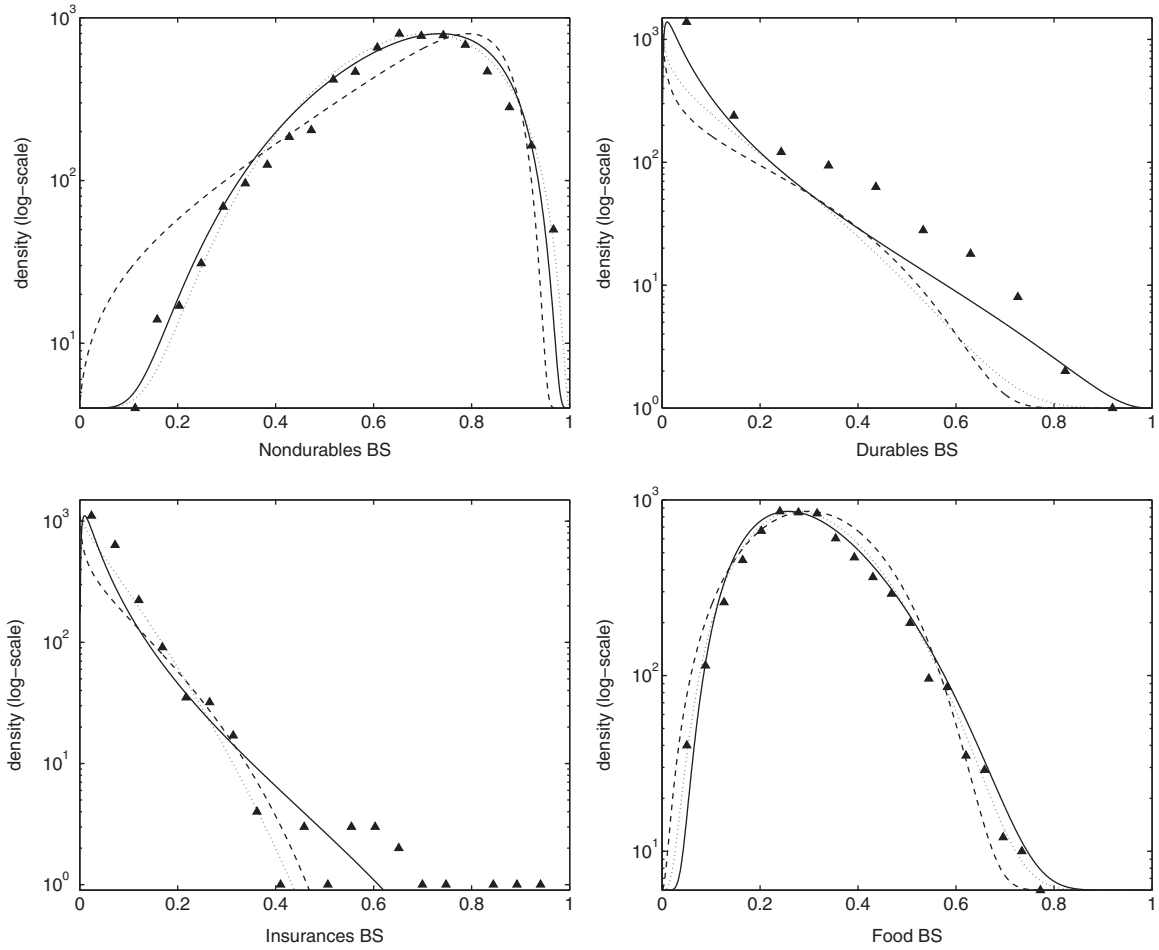

Fig. 12. Fitting alternative densities to household budget share distributions. Wave: 2004. Beta, dotted line; LN-B, solid line; InvΓ-B, dashed line.

and estimated sample moments of household budget share distributions. Indeed, suppose to classify now commodity categories on the base of estimated sample moments only (i.e., without fitting household budget share distributions with any parametric model). In particular, suppose to focus on estimates of the mean (μ), the median (med), standard deviation (σ), skewness (ξ) and kurtosis (κ). Let us take the number of observations outside the estimated interval $[\mu-\sigma, \mu+\sigma]$ as a measure of dispersion of household budget share distributions: the larger this number the higher the dispersion around the mean. Let us also say that a household budget share distribution has low (high) mean if the latter is lower (higher) than the median. Finally, let us discriminate between left-skewed ($\xi<0$) and right-skewed ($\xi>0$) distributions; and call a distribution fat-tailed if $\kappa>>3$.

Given this setup, one gets the two taxonomies of Table 10. Notice first that apart from the position of nondurables in the right taxonomy (the one involving kurtosis and standard deviation), both taxonomies reproduce the ones obtained using estimated parameters. More specifically, durables and insurance budget share distributions have mean lower than the median ($\mu/med<1$), low dispersion, they are highly right-skewed ($\xi>0$) and fat-tailed ($\kappa>>3$). Nondurable budget share distributions display instead a mean similar to the median ($\mu/med\simeq 1$), are left-skewed ($\xi<0$), and have thinner, but still thicker than a normal, tails ($\kappa\geq 3$). This taxonomy has a rather interesting economic meaning, somewhat related with Engel's classification of commodities. [7] Indeed, we do not expect many extreme observations, and therefore a higher kurtosis, when dealing with the consumption of nondurable goods (more likely to be related to necessary goods), while exceptional events are more common when dealing with durable goods (category which includes luxury goods).

This simple exercise has one main implication. It shows that the proposed density family, in addition to its other appealing properties, can be easily employed – via the evaluation of the estimated parameters – to build classifications of commodity categories, which are also consistent with other taxonomies developed on the basis of estimated sample moments. In our view, the classification built

[7] Notice however that whereas Engel's classification is based on the concept of necessity, the taxonomy introduced here stems from an empirical analysis of consumption expenditures. On this point see Chai and Moneta (2010).

Table 8
Average absolute deviation statistics with p-values (in brackets). Average absolute deviations computed by comparing empirical and theoretical pdfs. N, nondurables; D, durables; I, insurances; F, food. In boldface the figures that in any given wave and commodity category minimize average absolute deviations or maximize p-values.

	pdf	Waves							
		1989	1991	1993	1995	1998	2000	2002	2004
N	Beta	0.004 (0.44)	0.005 (0.86)	0.008 (0.70)	0.010 (**0.26**)	0.010 (**0.46**)	0.015 (0.19)	0.013 (**0.59**)	0.014 (**0.64**)
	LN-B	0.003 (0.43)	0.004 (0.86)	0.007 (0.68)	0.009 (0.25)	0.010 (**0.46**)	0.015 (**0.20**)	0.013 (0.58)	0.014 (**0.64**)
	InvΓ-B	**0.00** (**0.66**)	**0.003** (**0.91**)	**0.006** (**0.73**)	**0.008** (0.25)	**0.008** (0.45)	**0.013** (**0.20**)	**0.011** (0.58)	**0.012** (0.62)
D	Beta	0.004 (0.88)	0.003 (0.94)	0.005 (0.68)	0.004 (0.77)	0.004 (0.85)	0.004 (0.79)	0.005 (0.66)	0.005 (0.66)
	LN-B	**0.002** (**0.93**)	**0.002** (**0.99**)	**0.004** (0.57)	**0.003** (**0.87**)	**0.002** (**0.97**)	**0.003** (**0.86**)	**0.004** (**0.70**)	**0.003** (**0.74**)
	InvΓ-B	0.004 (0.90)	0.003 (0.96)	0.005 (**0.71**)	0.004 (0.82)	0.004 (0.87)	0.004 (0.78)	0.006 (0.65)	0.006 (0.62)
I	Beta	0.003 (0.77)	0.004 (0.63)	**0.002** (**0.95**)	0.003 (0.71)	0.002 (0.94)	**0.003** (0.76)	0.003 (0.62)	**0.002** (0.76)
	LN-B	**0.002** (**0.97**)	**0.003** (**0.89**)	0.003 (0.87)	**0.002** (**0.83**)	0.003 (0.81)	**0.003** (**0.77**)	**0.002** (**0.92**)	**0.002** (**0.95**)
	InvΓ-B	0.003 (0.76)	0.004 (0.64)	**0.002** (0.94)	0.003 (0.69)	**0.001** (**0.97**)	**0.003** (**0.77**)	**0.002** (0.68)	0.003 (0.54)
F	Beta	**0.001** (**1.00**)	**0.002** (**0.99**)	**0.004** (**0.78**)	**0.006** (0.61)	**0.004** (**0.80**)	**0.005** (0.74)	**0.005** (**0.80**)	**0.007** (**0.53**)
	LN-B	**0.001** (**1.00**)	**0.002** (0.98)	**0.004** (**0.78**)	**0.006** (**0.62**)	**0.004** (0.79)	**0.005** (**0.75**)	**0.005** (0.79)	**0.007** (0.52)
	InvΓ-B	0.002 (**1.00**)	0.003 (0.98)	0.005 (0.77)	0.007 (0.60)	0.005 (0.77)	0.006 (0.73)	0.006 (**0.80**)	0.008 (0.52)

Table 9
A taxonomy of household budget share distributions according to the estimated parameters (m, s) and (p, θ). N, nondurables; D, durables; I, insurances; F, food.

LN-B	$m>0$	$m<0$	InvΓ-B	$\theta>1$	$\theta<1$
$s<1$	F	N	$p>1$	F	N
$s>1$	D, I		$p<1$	D, I	

Table 10
A taxonomy of household budget share distributions according to estimated sample moments. N, nondurables; D, durables; I, Insurances; F, food. μ, mean; *med*, median; σ, standard deviation; ξ, skewness; κ, kurtosis.

	$\xi>0$	$\xi<0$	$\kappa \gg 3$ Low σ	$\kappa \geq 3$ High σ
$\mu/med \simeq 1$	F	N		N, F
$\mu/med<1$	D, I		D, I	

using estimated parameters of LN-B and InvΓ-B densities (Table 9) should be preferred to the one based on sample moments (Table 10) for at least two reasons. First, it is more parsimonious, as it entails the estimation of only two parameters. Second, it is obtained through a statistically sound parametric model of the whole household budget share distribution, and hence – unlike that based on sample moments – is based on a full description of the sample.

8. Conclusions

In this paper we have explored the statistical properties of household consumption expenditure and budget share distributions for a large sample of Italian households in the period 1989–2004. In a previous paper (Fagiolo et al., 2010), we have studied the statistical properties of (unconditional and age-conditioned) Italian household consumption expenditure (HCE) distributions. Here, we tackle the issue of exploring the statistical properties of (unconditional) budget share distributions. The starting point is the observation that HCE distributions – disaggregated over consumption categories – are not statistically independent and that makes very hard to predict the shape of BS distributions, even if we knew how HCE marginals were distributed.

A preliminary descriptive analysis has shown that the shapes of such distributions are relatively stable across time but display a lot of across-commodity heterogeneity. In addition, multivariate analyses have suggested that the most-widely used parametric models for data on the simplex are not very successful in describing the data and, in any case, do not allow for interesting and straightforward economic interpretations. Therefore, we have turned to a univariate analysis. We have derived a family of parsimonious parametric models (densities) for univariate household (not transformed) budget-share distributions that are consistent with the statistical properties of observed household consumption expenditure distributions (which household budget share distributions are computed from) and are able to satisfactorily fit the observed data while accommodating the existing shape heterogeneity. Finally, we have shown that the estimated parameters of such densities can be employed to build

economically meaningful taxonomies of commodity categories, which partly map into the standard classification of goods into necessary, luxury or inferior.

Given its preliminary nature, the present work allows for many possible extensions. First, the foregoing exercises can be replicated on similar databases of other countries, possibly at different levels of commodity category disaggregation. This may help in assessing the robustness and generality of our findings. In particular, sensitivity analyses where one compares distributional analyses at different aggregation levels, may shed some light on whether our results are generalizable beyond Italian data and reflect true empirical regularities household expenditure data irrespective of the data source and disaggregation level employed.

Second, as already discussed in Section 2, one may consider to link more closely the approach pursued here with that employed in Engel-curve-related works (Lewbel, 2008). More specifically, instead of focusing only on unconditional budget share distributions, one might think to study the shape (and the moments) of household budget share distributions conditional to household income or total expenditures, age and cohort of household's head, and other relevant household- or commodity-specific variables. The idea here is to go beyond standard parametric or non-parametric Engel-curve studies and look not only at how the first (and maybe second) moment of such conditional distributions changes with household income or total expenditure, but also at how the whole shape of conditional household budget share distributions is affected by increasing income levels (and across different commodity categories).

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

Aitchison, J., Egozcue, J., 2005. Compositional data analysis: where are we and where should we be heading? Mathematical Geology 37, 829–850.

Anderson, T.W., Darling, D.A., 1954. A test for goodness-of-fit. Journal of the American Statistical Association 49, 765–769.

Attanasio, O., 1999. Consumption demand. In: Taylor, J., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier Science, Amsterdam.

Axtell, R., 2001. Zipf distributions of U.S. firm sizes. Science 293, 1818–1820.

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society B 61, 579–602.

Azzalini, A., Dalla Valle, A., 1996. The multivariate skew-normal distribution. Biometrika 83, 715–726.

Banks, J., Blundell, R., Lewbel, A., 1997. Quadratic Engel curves and consumer demand. The Review of Economics and Statistics 79, 527–539.

Battistin, E., Blundell, R., Lewbel, A., 2007. Why is Consumption More Log Normal than Income? Gibrat's Law Revisited. IFS Working Papers WP08/07. The Institute for Fiscal Studies, London, U.K.

Battistin, E., Miniaci, R., Weber, G., 2003. What can we learn from recall consumption data? Journal of Human Resources 38, 354–385.

Beaulieu, N.C., Abu-Dayya, A.A., McLane, P.J., 1995. Estimating the distribution of a sum of independent lognormal random variables. IEEE Transactions on Communications 43, 2869–2873.

Bentler, P., Bonett, D., 1980. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88, 588–606.

Bera, A., Jarque, C., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6, 255–259.

Bera, A., Jarque, C., 1981. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. Economics Letters 7, 313–318.

Blundell, R., 1988. Consumer behaviour: theory and empirical evidence – a survey. The Economic Journal 98, 16–65.

Blundell, R., Chen, X., Kristensen, D., 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. Econometrica 75, 1613–1699.

Bottazzi, G., Secchi, A., 2006. Explaining the distribution of firm growth rates. RAND Journal of Economics 37, 235–256.

Brandolini, A., 1999. The Distribution of Personal Income in Post-war Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality. Temi di Discussione (Economics Working Papers), No. 350, Bank of Italy, Rome.

Caselli, F., Ventura, J., 2000. A representative consumer theory of distribution. American Economic Review 90, 909–926.

Chai, A., Moneta, A., 2008. Satiation, Escaping Satiation, and Structural Change. Some Evidence from the Evolution of Engel Curves. Max Planck Institute for Economics, Jena, Germany.

Chai, A., Moneta, A., 2010. Retrospectives: Engel curves. Journal of Economic Perspectives 24, 225–240.

Chatterjee, A., Yarlagadda, S., Chakrabarti, B. (Eds.), 2005. Econophysics of Wealth Distributions. Springer-Verlag Italia, Milan.

Clementi, F., Gallegati, M., 2005. Pareto's law of income distribution: evidence for Germany, the United Kingdom, and the United States. In: Chatterjee, A., Yarlagadda, S., Chakrabarti, B. (Eds.), Econophysics of Wealth Distributions. Springer-Verlag Italia, Milan.

Connor, R.J., 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association 64, 194–206.

D'Agostino, R., 1970. Transformation to normality of the null distribution of G1. Biometrika 57, 679–681.

D'Agostino, R., Stephens, M., 1986. Goodness of Fit Techniques. Marcel Dekker, New York.

Deaton, A., 1992. Understanding Consumption. Clarendon Press, Oxford.

Engel, E., 1857. Die Produktions und Consumptions Verhaltnisse des Konigreichs Sachsen. Zeitschrift des Statistisehen Bureaus des Koniglich Sachsischen Ministerium des Innern, 22.

Engel, J., Kneip, A., 1996. Recent approaches to estimating Engel curves. Journal of Economics 63, 187–212.

Evans, M., Hastings, N., Peacock, B., 2000. Statistical Distributions, 3rd ed. Wiley, New York.

Fagiolo, G., Alessi, L., Barigozzi, M., Capasso, M., 2010. On the distributional properties of household consumption expenditures. The case of Italy. Empirical Economics 38, 717–741.

Fagiolo, G., Napoletano, M., Roventini, A., 2008. Are output growth-rate distributions fat-tailed? Some evidence from OECD countries. Journal of Applied Econometrics 23, 639–669.

Fenton, L.F., 1960. The sum of lognormal probability distributions in scatter transmission systems. IRE Transactions on Communication System CS-8, 57–67.

Forni, M., Lippi, M., 1997. Aggregation and the Microfoundations of Dynamic Macroeconomics. Oxford University Press, Oxford.

Fry, J.M., Fry, T.R., McLaren, K.R., 1996. The stochastic specification of demand share equations: restricting budget shares to the unit simplex. Journal of Econometrics 73, 377–385.

Gallegati, M., Kirman, A.P. (Eds.), 1999. Beyond the Representative Agent. Aldershot and Lyme, Edward Elgar.

Groeneveld, R., Meeden, G., 1984. Measuring skewness and kurtosis. The Statistician 33, 391–399.

Hartley, J.E., 1997. The Representative Agent in Macroeconomics. Routledge, London, New York.

Hendry, D.F., 2000. Econometrics: Alchemy Or Science?: Essays in Econometric Methodology. Oxford University Press, Oxford.

Hildenbrand, W., 1994. Market Demand: Theory and Empirical Evidence. Princeton University Press, Princeton.

Huber, P., 1981. Robust Statistics. John Wiley and Sons, New York.

Ibragimov, R., 2005. On the Robustness of Economic Models to Heavy-taildness Assumptions, Bundesbank, November 2005 Conference.

Kaldor, N., 1961. Capital accumulation and economic growth. In: Lutz, F., Hague, D. (Eds.), The Theory of Capital. St. Martin's Press, New York.

Kirman, A.P., 1992. Whom or what does the representative individual represent? Journal of Economic Perspectives 6, 117–136.

Kuiper, N.H., 1962. Tests concerning random points on a circle. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 63, 38–47.

Lewbel, A., 2008. Engel curves. In: Durlauf, S.N., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics. , 2nd ed. Palgrave Macmillan.

Lilliefors, H., 1967. On the Kolmogorov–Smirnov tests for normality with mean and variance unknown. Journal of the American Statistical Association 62, 399–402.

Mateu-Figueras, G., Pawlowsky-Glahn, V., 2007. The skew-normal distribution on the simplex. Communications in Statistics–Theory and Methods 36, 1787–1802.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Barceló-Vidal, C., 2007a. The additive logistic skew-normal distribution on the simplex. Stochastic Environmental Research and Risk Assessment 36,205–214.
Mateu-Figueras, G., Puig, P., Pewsey, A., 2007b. Goodness-of-fit tests for the skew-normal distribution when the parameters are estimated from the data. Communications in Statistics–Theory and Methods 36, 1735–1755.
McLaren, K.R., Fry, J.M., Fry, T.R.L., 1995. A simple nested test of the almost ideal demand system. Empirical Economics 20, 149–161.
Mehta, N.B., Wu, J., Zhang, J., 2006. Approximating the sum of correlated lognormal or lognormal-Rice random variables. IEEE International Conference on Communications (ICC) 4, 1605–1610.
Milevski, M.A., Posner, S.E., 1998. Asian options, the sum of log-normals, and the reciprocal gamma distribution. Journal of Financial and Quantitative Analysis 33, 409–422.
Moors, J., 1988. A quantile alternative to kurtosis. The Statistician 37, 25–32.
Pasinetti, L.L., 1981. Structural Change and Economic Growth. Cambridge University Press, Cambridge.
Pearson, E.S., Stephens, M.A., 1962. The goodness-of-fit tests based on Wk and Uk. Biometrika 49, 397–402.
Prais, S.J., Houthakker, H.S., 1955. The Analysis of Family Budgets. Cambridge University Press, Cambridge.
Safak, A., Safak, M., 1994. Moments of the sum of correlated lognormal random variables, IEEE 44th Vehicular Technology Conference, vol. 1, 140–144.
Schwartz, S., Yeh, Y., 1982. On the distribution function and moments of power sums with lognormal components. The Bell System Technical Journal 61, 1441–1462.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society B 53, 683–690.
Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Szekely, G., Rizzo, M., 2005. A new test for multivariate normality. Journal of Multivariate Analysis 93, 58–80.