PAPERS on Economics & Evolution

1101

The Rank of a System of Engel Curves.
How Many Common Factors?

by

Matteo Barigozzi
Alessio Moneta

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: evopapers@econ.mpg.de

ISSN 1430-4716

Max Planck Institute of Economics
Evolutionary Economics Group
Kahlaische Str. 10
07745 Jena, Germany
Fax: ++49-3641-686868

THE RANK OF A SYSTEM OF ENGEL CURVES. HOW MANY COMMON FACTORS?

Matteo BARIGOZZI[1] Alessio MONETA[2]

January 27, 2011

Abstract

By representing a system of budget shares as an approximate factor model we determine its rank, i.e. the number of common functional forms, or factors, spanning the space of Engel curves. Once the common factors are estimated via approximate principal components, we identify them by imposing statistical independence. Finally, by means of parametric and non-parametric regressions we estimate the factors as functions of total expenditure. Using data from the U.K. Consumption Expenditure Survey from 1968 to 2006, we find evidence of three common functional forms which correspond to decreasing, increasing and almost constant Engel curves. The household consumption behavior is therefore driven by three factors respectively related to necessities (e.g. food), luxuries (e.g. vehicles), and goods to which is allocated the same percentage of total budget both in rich and in poor households (e.g. housing).

Keywords: Engel Curves, Demand Systems, Factor Models, Independent Component Analysis.

JEL classification: C52, D12.

[1]Department of Statistics, London School of Economics and Political Science, U.K. Email: *m.barigozzi@lse.ac.uk*

[2]*Corresponding author.* Max Planck Institute of Economics, Kahlaische Strasse 10, 07745 Jena, Germany. Email: *moneta@econ.mpg.de*

We are grateful to Andreas Chai, Giorgio Fagiolo, and Ulrich Witt for comments on an earlier version of the paper. We thank Stefanie Picard for research assistance. We also thank the U.K. Central Statistical Office for making available the U.K. Family Expenditure Survey and the Expenditure and Food Survey data through the Economic and Social Data Service. We retain responsibility for any error.

1 Introduction

We investigate the problem of estimating the rank of a system of Engel curves by using factor analysis. Engel curves represent the dependence of different categories of expenditures on total income, usually proxied by total expenditure (Banks et al., 1997). We consider here Engel curves expressed in the budget share form as $w_j = h_j(x)$, where w_j is the proportion of total expenditure x on the category of expenditure j. Following the literature (see e.g. Lewbel, 1991), we assume that h_j is the sum of different functional forms: $h_j = \sum_r a_{rj} g_r(x)$. The rank of a system of Engel curves is defined as the maximum dimension of the function space spanned by h_j, i.e. the rank of the matrix formed by the elements a_{rj}. The estimation of the rank of this system has concerned much literature on empirical analysis of consumption (see Gorman, 1981; Lewbel, 1991; Kneip, 1994; Donald, 1997; Banks et al., 1997, among others). It has indeed been shown that the rank has several logical connections with the properties of consumer preferences, separability and aggregation of demands (Lewbel, 1991). The most remarkable result is the proof in Gorman (1981) that if consumers are utility maximizer agents, then the rank of the demand system has to be three at maximum. However, the reverse does not hold: if the rank of demand system is three, this does not necessarily mean that the underlying preferences have a particular structure or that consumption behaviors match up with those postulated by the model of rational choice (cfr. Aversi et al., 1999)

We formalize the system of Engel curves as a factor model and show that finding the maximum rank amounts to estimating the maximum number of common factors as suggested in Bai and Ng (2002). We apply the model to U.K. Family Expenditure Survey annual data from 1968 to 2006. Factor models are usually applied to panels of time series under the assumption of large cross-section and time dimension (see e.g. Stock and Watson, 1989; Forni et al., 2000; Bai and Ng, 2002, among others). In this paper the cross-section dimension is made of a large number of budget shares relative to 13 different goods pooled over few years, while the time dimension is substituted by 100, income determined, representative households (the database is built exactly as in Kneip, 1994). Exploiting this large panel we are able to identify the number of factors by eschewing any assumption of cross-sectional uncorrelation among idiosyncratic shocks. The method employs the criterion for determining the number of factors proposed by Bai and Ng (2002) and recently refined by Alessi et al. (2010).

In the majority of the panels considered we find evidence of a maximum of three common factors. The variance explained by the common factors constantly decreased in time since the early 1970s. Once the factors are estimated via approximate principal components (see Bai and Ng, 2002), we exploit their non-Gaussianity to achieve identification by imposing statistical independence (see Hyvärinen et al., 2001, for a review on the possible algorithms used for estimation). Finally, by means of different regressions of the identified factors on total expenditure, we find the common functional forms of the system of budget shares Engel curves. This last step is accomplished both in a parametric way (choosing the same regressors as Lewbel, 1991) and non-parametrically (using kernel regressions). The household consumption behavior turns out to be driven by at most three different functions of total expenditure corresponding to the standard classification of goods: *i*) a decreasing function capturing consumption necessities (e.g. food), *ii*) an

increasing function related to luxuries (e.g. vehicles), and *iii*) an almost constant function corresponding to the expenditure for goods to which is allocated the same percentage of total budget both in rich and in poor households (e.g. housing).

In section 2 we show the economic implications of different values of the rank of a system of Engel curves. In section 3 we represent the system as an approximate factor model, we explain the approximate principal components estimation method, the related test for the number of common factors, and identification via independent component analysis. In section 4 we describe the data used and the way we built the dataset used. In section 5 we show results on the number of factors and their interpretation as non-linear functions of total expenditure. In section 6 we conclude.

2 Theoretical framework

In this paper we study the properties of a system of Engel curves of the form:

$$w_{jh} = \sum_{r=1}^{R} a_{jr} g_r(x_h), \quad j = 1, \ldots, J, \; h = 1, \ldots, H \tag{1}$$

where w_{jh} is the budget share of household h spent on good j , x_h is total consumption expenditure, and we assume $R \leq J$. Or in vector notation:

$$\mathbf{w}_h = \mathbf{A}\, \mathbf{g}(x_h), \quad h = 1, \ldots, H, \tag{2}$$

where $\mathbf{w}_h$, $\mathbf{A}$, and $\mathbf{g}(x_h)$ have dimensions $J \times 1$, $J \times R$, and $R \times 1$ respectively. Since $R \leq J$, R is the rank of the matrix $\mathbf{A}$ and determines the maximum dimension of the function space spanned by Engel curves. Gorman (1981) and Lewbel (1991) prove that the knowledge of R provides important implications about the functional form, separability, and aggregability of consumer preferences. In particular, Lewbel (1991) shows that:[1]

(i) if $R = 1$, and the *adding-up* condition holds, then budget shares are constant across income. Indeed, the adding up conditions requires that $\sum_{j=1}^{J} w_j = 1$, J being the total number of goods in which the budget is subdivided. Thus, we have that $\sum_{j=1}^{J} a_{1j} g_1(x_h) = 1$. Hence, $g_1(x_h) = (\sum_{j=1}^{J} a_{1j})^{-1}$ which is a constant. Therefore, since each budget share is $w_{jh} = a_{1j} g_1(x_h)$, this implies that actually w_{jh} does not depend on x_h;

(ii) if $R = 2$, then the underlying demand functions are generalized linear. The so-called AIDS, trans-log, linear expenditure, PIGL, and PIGLOG models are all rank-two models;

[1] We have to notice that although Lewbel (1991) considers the model using the logarithms of total expenditure, while we specify it using total expenditure as explanatory variable, the economic implications of the model conclusions do not change. Indeed, by assuming that g_r can be non-linear functions of total expenditure we implicitly allow for the log dependence.

(iii) if the system of equations (1) is an *exactly aggregable* class of demand, that is the aggregate (across households) demand depend only on the means of the individual demands q_h, then utility maximization constrains the maximum number of R to three (Gorman, 1981).

Concerning the last case, Aversi et al. (1999) simulate micro-founded models of consumption expenditure which indirectly support Gorman's rank-three assumption, independently of the level of aggregation over goods. This, however, happens despite the fact that the simulated individual behaviors are designed by the author to be at odds with those postulated by the standard utility-based model of rational choice. Therefore, we can conclude that $R \leq 3$ is a necessary but not sufficient condition for having utility-maximizer consumers.

3 Econometric setup

3.1 An approximate factor model for budget shares

Following Lewbel (1991), we consider equation (1) with a noise term added:

$$w_{jh} = \sum_{r=1}^{R} a_{jr} g_r(x_h) + e_{jh}, \quad j = 1, \ldots, J, \; h = 1, \ldots, H, \tag{3}$$

This is a factor model with R factors which are common to the J budget shares, where we assume $R < J$. We call the first term on the right hand side the common component and the second the idiosyncratic component, which by assumption has zero mean. An *exact* factor model would require that idiosyncratic components are uncorrelated across goods, i.e. for any household h we would need $\mathrm{E}[e_{ih}e_{jh}] = 0$ for any $i \neq j$. This is an unreasonable restriction. Indeed, the whole budget must sum to one, i.e. $\sum_{j=1}^{J} w_{jh} = 1$, for every household h, which implies non zero correlation across goods both in the common and in the idiosyncratic component. However, if J is large, we can allow for mildly correlated idiosyncratic terms. A large cross-section of budget shares allows us to choose a different modeling and estimation strategy with respect to Lewbel (1991). Namely, we can represent budget shares with an *approximate* factor structure as in Bai and Ng (2002).

For every household h, (3) can be written using vector notation as

$$\mathbf{w}_h = \mathbf{A}\mathbf{f}_h + \mathbf{e}_h, \qquad h = 1, \ldots, H, \tag{4}$$

where $\mathbf{w}_h$ is a J-dimensional zero-mean vector, $\mathbf{A}$ is the $J \times R$ loadings matrix, and $\mathbf{f}_h$ is an R-dimensional vector of latent factors (independent of j) and driving the common component of each of the J budget shares w_{jh}. Finally, $\mathbf{e}_h$ is a zero mean J-dimensional vector of possibly mildly correlated idiosyncratic noises.

We recall the main Assumptions of the model by Bai and Ng (2002):[2]

1. Factors:
   (a) $\mathrm{E}[f_{ih}] = 0$, for any $i = 1, \ldots, R$ and any $h = 1, \ldots, H$,
   (b) $\lim_{H\to\infty} \frac{1}{H} \sum_{h=1}^{H} \mathbf{f}_h \mathbf{f}_h' = \mathbf{\Sigma}^{\mathbf{f}}$, for some positive definite and diagonal $R \times R$ matrix $\mathbf{\Sigma}^{\mathbf{f}}$,
   (c) $\mathrm{E}[||\mathbf{f}_h||^4] < \infty$;
2. Loadings:
   (a) $|a_{jr}| \leq \bar{a} < \infty$, for any $j = 1, \ldots, J$ and $r = 1, \ldots, R$,
   (b) $\lim_{J\to\infty} ||\mathbf{A}'\mathbf{A}/J - \mathbf{D}|| = 0$, for some positive definite $R \times R$ matrix $\mathbf{D}$;
3. Idiosyncratic components:
   (a) $\mathrm{E}[e_{jh}] = 0$, for any $j = 1, \ldots, J$ and any $h = 1, \ldots, H$,
   (b) define $\mathbf{\Sigma}^{\mathbf{e}} = \mathrm{E}[\mathbf{e}_h \mathbf{e}_h']$ then there exists $M > 0$ s.t. $\sum_{k=1}^{J} |(\mathbf{\Sigma}^{\mathbf{e}})_{jk}| \leq M$ for any $j = 1, \ldots, J$.

Assumption 1 implies the existence of the covariance matrix of the factors which we also assume to be uncorrelated. Assumption 2 is necessary for identification of the loadings and implies that, when J goes to infinity, $\mathbf{A}'\mathbf{A} = O(J)$. Assumption 3 defines an approximate factor model by allowing for some correlation across goods in the idiosyncratic components, this is equivalent to require boundedness of the largest eigenvalue of $\mathbf{\Sigma}^{\mathbf{e}}$ as J goes to infinity (see Chamberlain and Rothschild, 1983).

The rank of the considered system of budget shares, holding prices fixed, is therefore the smallest integer R such that (4) holds and where the R factors, common across goods, are non-linear functions of total expenditure x_h, i.e. $\mathbf{f}_h = [g_1(x_h) \ldots g_R(x_h)]'$. While Lewbel (1991) proposes a test based on LDU decomposition to determine R even in presence of small cross sections, both Kneip (1994) and Donald (1997) propose non-parametric estimation methods of both R and the functions g_r. We instead adopt here the approach by Bai and Ng (2002) who propose an estimation method based on approximate principal component analysis (PCA). This approach provides a consistent estimate of R and of the space spanned by the factors for both H and J going to infinity, without the need for non-parametric estimation of the g_r functions. Once the rank R of $\mathbf{w}_h$ is determined, the non-parametric functions evaluated at x_h are consistently estimated as the principal components, i.e. as the factors (see section 3.2). If $R > 1$, the functions are determined only up to an orthogonal transformation. We provide below an identification strategy based on statistical independence of the factors (see section 3.3). Finally, the functions g_r may be recovered from the identified factors via non-linear regression or non-parametric estimation.

[2] We define the norm of a generic matrix $\mathbf{B}$ as $||\mathbf{B}|| = \sqrt{\mathrm{Tr}(\mathbf{B}'\mathbf{B})}$.

3.2 Estimation of the factors

We follow Bai and Ng (2002) for the estimation of model (4) and of R. Let us collect all the budget shares into a $J \times H$ matrix $\mathbf{w} = (\mathbf{w}_1 \dots \mathbf{w}_H)$, and the factors into a $R \times H$ matrix $\mathbf{F} = (\mathbf{f}_1 \dots \mathbf{f}_H)$. First, let us assume that R and $\mathbf{F}$ are known, then the estimated loadings must satisfy

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} \mathrm{E}[(\mathbf{w} - \mathbf{A}\mathbf{F})(\mathbf{w} - \mathbf{A}\mathbf{F})'], \tag{5}$$

which is equivalent to minimizing the variance of the idiosyncratic component. In order to solve the above minimization, we need to impose an additional identification condition on the estimated loadings. Consistently with Assumption 2, we require $\widehat{\mathbf{A}}'\widehat{\mathbf{A}}/J = \mathbf{I}_R$, where $\mathbf{I}_R$ is the R-dimensional identity matrix. In this case the columns of $\widehat{\mathbf{A}}$ are given by $\sqrt{J}$-times the eigenvectors corresponding to the R largest eigenvalues of the covariance matrix of $\mathbf{w}$. Equivalently, we can look for maximum variance weighted averages of data, the weights being such that

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{A}} \mathrm{E}[(\mathbf{A}'\mathbf{w})(\mathbf{A}'\mathbf{w})'] \tag{6}$$

with the same identification condition as before, i.e. $\widehat{\mathbf{A}}'\widehat{\mathbf{A}}/J = \mathbf{I}_R$. The first order conditions derived from (5) and (6) are identical and therefore have the same solution for $\widehat{\mathbf{A}}$. Geometrically, we are just looking in a J-dimensional space for the R mutually orthogonal directions along which the variance of the observed data is maximum. Such directions are given by the normalized eigenvectors corresponding to the R largest eigenvalues of the covariance matrix of observed data.

From classical PCA, we know that if H is large the covariance matrix of $\mathbf{w}$ is estimated consistently as $\frac{1}{H}\sum_{h=1}^{H} \mathbf{w}_h \mathbf{w}_h'$, and therefore also its eigenvectors (i.e. the loadings) are estimated consistently. If J is large too, then also the factors can be consistently estimated as the R largest principal components: $\widehat{\mathbf{F}} = \widehat{\mathbf{A}}'\mathbf{w}/J$ (see Theorem 1 in Bai and Ng, 2002, for a proof).

To understand why in approximate factor models we can rule out the idiosyncratic components as J goes to infinity, we show that, in this case, all the relevant information contained in the data is summarized by the factors, which are nothing else but a weighted average of the data. Since by Assumptions 1 and 3, both the factors and the idiosyncratic components have zero mean, the covariance of $\widehat{\mathbf{A}}'\mathbf{w}/J$ is estimated as

$$\frac{1}{J^2}\widehat{\mathbf{A}}'\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{w}_h\mathbf{w}_h'\right)\widehat{\mathbf{A}} = \frac{1}{J^2}\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{f}_h\mathbf{f}_h'\right)\widehat{\mathbf{A}}'\widehat{\mathbf{A}} + \frac{1}{J^2}\widehat{\mathbf{A}}'\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{e}_h\mathbf{e}_h'\right)\widehat{\mathbf{A}}.$$

By defining $\mathbf{\Sigma}^{\mathbf{w}} = \mathrm{E}[\mathbf{w}_h\mathbf{w}_h']$ and given Assumptions 1 and 3, as H goes to infinity the previous expression converges in probability to

$$\frac{1}{J^2}\mathbf{A}'\mathbf{\Sigma}^{\mathbf{w}}\mathbf{A} = \frac{1}{J^2}\mathbf{A}'\mathbf{A}\mathbf{\Sigma}^{\mathbf{F}}\mathbf{A}'\mathbf{A} + \frac{1}{J^2}\mathbf{A}'\mathbf{\Sigma}^{\mathbf{e}}\mathbf{A} \leq \frac{1}{J^2}\mathbf{A}'\mathbf{A}\mathbf{\Sigma}^{\mathbf{F}}\mathbf{A}'\mathbf{A} + \frac{1}{J^2}M\mathbf{A}'\mathbf{A},$$

where the inequality and M are due to Assumption 3. Finally, from Assumption 2 we

know that $\mathbf{A}'\mathbf{A} = O(J)$, therefore as J goes to infinity, the contribute of the common component to the total variance of $\mathbf{A}'\mathbf{w}/J$ is $O(1)$, while the contribute of the idiosyncratic component becomes negligible being $O(J^{-1})$. This means that all the relevant information contained in $\mathbf{w}$ can be recovered by estimating the first R factors as weighted averages of $\mathbf{w}$ with weights given by the estimated eigenvectors $\widehat{\mathbf{A}}$ divided by the cross-sectional dimension J.

3.3 Estimation of the number of factors

Following Bai and Ng (2002), we can use the above estimation method to estimate the number of factors R. This can be done by estimating the factors and their loadings for different values k of the number of factors and by computing each time the average variance of the idiosyncratic component which is given in (5), call it $V(k, \widehat{\mathbf{A}}^k, \widehat{\mathbf{F}}^k)$, where $\widehat{\mathbf{A}}^k$ and $\widehat{\mathbf{F}}^k$ are estimates of loadings and factors when assuming the existence of k common factors. The true number of factors is the value of k that minimizes this function, conveniently penalized with a penalty function $p(k, J, H)$ that depends both on J and on H. In this paper we look for minima of the ICs criteria proposed by Bai and Ng (2002), i.e.

$$\widehat{R} = \underset{1 \leq k \leq k_{\max}}{\operatorname{argmin}} \log V(k, \widehat{\mathbf{A}}^k, \widehat{\mathbf{F}}^k) + p(k, J, H) \tag{7}$$

where

$$\begin{aligned} p(k, J, H) &= k \left(\frac{J+H}{JH} \right) \log \left(\frac{JH}{J+H} \right) \\ &\text{or} \\ p(k, J, H) &= k \left(\frac{J+H}{JH} \right) \log \left(\min \left\{ \sqrt{J}, \sqrt{H} \right\} \right)^2 . \end{aligned} \tag{8}$$

Provided that we have a consistent estimate of the factors and their loadings, Bai and Ng (2002) prove consistency of $\widehat{R}$ as J and H tend to infinity.

In the following sections we also apply a refinement of the information criteria by Bai and Ng (2002), proposed by Alessi et al. (2010) and the criterion by Onatski (2010) which is instead based on the asymptotic distribution of the eigenvalues of the sample covariance matrix.

3.4 Identification of the factors

Factor models have an indeterminacy which they cannot solve: both the estimated loading matrix $\widehat{\mathbf{A}}$ and factors $\widehat{\mathbf{F}}$ are asymptotically consistent estimates of the true ones only up to an orthogonal transformation. We have, therefore, an identification problem which makes difficult the economic interpretation of the estimated factors. In order to identify the model we use independent component analysis (ICA) which is based on the imposition of two further assumptions on the R latent factors:

4. the components of the factor vector f_{ih} are mutually independent, i.e. the joint cumulative distribution of the factors is given by

$$\mathcal{F}(\mathbf{f}_h) = \prod_{i=1}^{R} F_i(f_{ih}), \quad h = 1 \ldots, H,$$

   where F_i is the marginal cumulative distribution of the i-th factor;

5. the marginal distributions F_i are not Gaussian, for all $i = 1, \ldots, R$ with the exception of at most one.

Assumption 4 can be justified on the basis of the fact that the latent factors represent the common functional forms of a system of Engel curves. These forms, in turn, have characteristics which reflect fundamental aspects of human behaviors that drive consumption decisions. Consumption decisions can be seen as driven by basic needs and acquired wants. Therefore, assuming that latent factors are independent amounts to claim that the set of needs and wants associated with each factor is of fundamental different nature, i.e. generates an independent pattern, from the set of needs and wants associated with the other factors. For example, if a factor reflects a pattern associated with necessities and another factor reflects a pattern associated with luxuries, these two factors can be seen as statistical independent, because necessities mainly reflect physiological needs, while luxuries reflect culturally acquired wants such as social recognition and status. The drivers underlying consumption decisions about necessities and luxuries react in an independent way to changes in income: for example, physiological needs tend rapidly to satiate, as income gives the possibility to satisfy these needs, whereas acquired wants such as social recognition and status may be even increasingly reinforced, as income increases.

Assumption 5 can be justified by testing for normality in the data and also by noticing that often data on consumption expenditures are non-Gaussian (see e.g. Fagiolo et al., 2010) and, moreover, being budget shares defined on the unit interval, they must have a distribution with bounded support (e.g. a beta distribution) hence not a Gaussian distribution. A consequence of Assumption 5 is that also the joint distribution of the factors is not Gaussian.

ICA can been seen as an extension or a strengthening of PCA (see Comon, 1994; Hyvärinen et al., 2001; Bonhomme and Robin, 2009). Indeed, while PCA gives a transformation of the original space such that the computed latent factors are linearly uncorrelated, ICA goes further by attempting to minimize all statistical dependencies between the resulting components. One can show that *if* there exists a representation with non-Gaussian, statistically independent components then the representation is essentially unique (up to a permutation, a sign, and a scaling factor) (Comon, 1994). There exist a number of computationally efficient algorithms for consistent estimation (Hyvärinen et al., 2001).

The most popular ICA algorithms are: Joint Approximate Diagonalization of Eigenmatrices (JADE Cardoso and Souloumiac, 1993), Fast Fixed-Point Algorithm (Hyvärinen and Oja, 2000, FastICA). Both methods are based on two steps: *i*) a whitening step achieved by PCA, in which the data are transformed so that the covariance matrix is

diagonal and has reduced rank, i.e. we get rid of the idiosyncratic component; *ii*) a source separation step in which the orthogonal transformation necessary for achieving identification is determined.

When data usually tend to exhibit fat-tailed distributions and poor serial correlation (in our framework we have no correlation across households), JADE and FastICA, which are based on non-Gaussianity of the data, hence on higher order moments, are the most used algorithms. [3] We consider here only JADE, the results obtained with FastICA being similar.

Once estimation of the common component is accomplished via approximate PCA, we are left with a first estimate of the factors $\widehat{\mathbf{f}}_h$ for any household h. JADE looks for an orthogonal $J \times R$ matrix $\widehat{\mathbf{U}}$ such that the identified factors $\widetilde{\mathbf{f}}_h = \widehat{\mathbf{U}}'\widehat{\mathbf{f}}_h$ are maximally non-Gaussian distributed. A set of random vectors is mutually independent if all the cross-cumulants (i.e. the coefficients of the Taylor series expansion of the characteristic function) of order higher than two are equal to zero. In particular, Cardoso and Souloumiac (1993) prove that the factors $\widetilde{\mathbf{f}}_h$ are maximally independent if their associated fourth-order cumulant tensor which is a $R \times R$ matrix is maximally diagonal. Roughly speaking, if we call $\widehat{\mathbf{Q}}$ the estimated fourth-order cumulant tensor of $\widehat{\mathbf{f}}_h$ then $\widehat{\mathbf{U}}$ must be such that $\widehat{\mathbf{U}}'\widehat{\mathbf{Q}}\widehat{\mathbf{U}}$ is diagonal.[4] JADE is a very efficient algorithm in low dimensional problems as the one treated here (we have few factors), while a higher computational cost is required when the dimension increases.

Once we apply ICA the estimated and identified factors, $\widetilde{\mathbf{f}}_h$, are identified up to a permutation, a sign, and a scaling factor. The order of the factor is irrelevant for our purposes so it is left undetermined. Given that independent components are nothing else but weighted averages of the data, the sign is chosen to be consistent with the average of budget shares across goods. Finally, the scale is determined in such a way that the estimated factors have unit variance.

4 Data

The data set on which we estimate our model is built using data from the U.K. family expenditure survey (FES) 1968-2001 jointly with the expenditure and food survey (EFS) 2002-2006. We have data about household expenditures on various categories of goods and services. Each year approximately 7000 households were randomly selected, and each of them recorded expenditures for two weeks. We are able to recover information about total expenditures and expenditures on fourteen aggregated categories: (1) housing (net); (2) fuel, light, and power; (3) food; (4) alcoholic drinks; (5) tobacco; (6) clothing

[3] Another algorithm is Second-Order Blind Identification (SOBI Belouchrani et al., 1997), which, although usually applied in time-series analysis, could be extended to cross-sectional data with correlations among observations. However, this is not the case for us, as we assume no correlations across households.

[4] While the cumulant depends on four indexes the cumulant tensor depends on two indexes, the other two being canceled by means of an additional arbitrary matrix. We thus have to consider several cumulant matrices which have to be jointly diagonalized. We omit here the details, while giving just a general idea of the algorithm.

and footwear; (7) household goods; (8) household services; (9) personal goods and services; (10) motoring, fares and other travel; (11) leisure goods; (13) leisure services; and (14) miscellaneous and other goods. The fourteen categories add up to total expenditure. In order not to have to impose the adding-up condition on budget shares, we omit from our analysis the last category of expenditure and we restrict therefore to thirteen categories.[5]

We build the dataset used in the next sections as follows (see also Kneip, 1994, for a similar procedure). In order to have samples of households which are demographically homogeneous, we only consider families which have a number of members between two and three. Families of this type are approximately 3000 each year. We pool together budget shares over different years (we choose windows of 3, 5, and 10 years) in order to deal with a number of categories J big enough to estimate the model. Pooling together 10 years, for example, we are able to get $J = 13 \times 10 = 130$. To be able to pool together data over different years, we need to build normalized total expenditure. For each year considered, we divide data on total expenditure by the mean across households. Let X_{it} be the normalized total expenditure of household i in year t, then $W_{git} = X_{git}/X_{it}$ denotes the resulting budget share of household i, in year t, when the expenditure for good $g = 1, \dots 13$ is X_{git}. Starting from the original data, (X_{it}, W_{git}), we want to obtain a new set of pooled data (x_h, w_{jh}) for every j ranging over the 13 categories multiplied by the number of years under consideration, and h ranging over representative households as defined below. The pooling of budget shares consists in three steps. First, we specify a domain $X_{it} \in [0.25, 1.75]$. Since the value 1 for X_{it} corresponds to the average total expenditure in the year t, we exclude data for very rich and very poor households, which, being quite sparse, are not very reliable. Second, we specify a grid of equidistant values of total expenditure which do not depend on the chosen household nor on the year: $0.25 = \kappa_0 < \kappa_1 < \kappa_2 < \ldots < \kappa_H = 1.75$, with $H = 100$. Third, given a window $T = 3, 5, 10$ years, and for every $j = 1, \dots, 13 \cdot T$, we define a new set of data points (x_h, w_{jh}). The new budget share w_{jh} denotes the average across households of W_{git}, when considering only data for good g in years $t = 1, \dots T$, and for households i such that the corresponding total expenditure X_{it} is such that $X_{it} \in [(\kappa_{k-1}+\kappa_k)/2, (\kappa_k+\kappa_{k+1})/2]$. The normalized total expenditure for the representative household h is then $x_h = \kappa_h$. Finally, since $x_1, \dots, x_{100}$ are equidistant and normalized (scale free) values, in what follows for graphical convenience we let $x_1 = 1, x_2 = 2, \dots, x_{100} = 100$.

In table 1 we report the average (across households) budget shares for all 13 the considered goods and for the three 10-years windows considered. The majority of the budget is spent for food and housing. However, while the share of budget allocated to food has decreased since the early 1980s from 24% to 19%, the share allocated to housing has remained constant at an average level of 16%. A slight decrease is found also for fuel, light, and power budget shares from 7% to 4%. A smaller fraction of budget is allocated to all other goods and remained constant in time at values less than 10%. Two exceptions

[5]From 1987 to 2006 the survey contains a macro-code for each of the 13 categories. From 1968 to 1986 the FES contains macro-codes only for the first 6 categories (from housing to clothing and footwear), plus other macro-categories which are not consistent with the other 7 categories listed above (household goods, household services, personal goods and services, motoring, fares and other travel, leisure goods, and leisure services). We thus constructed, for the years 1968-1986, these 7 macro-categories aggregating micro-categories (disaggregate expenditures) in order that they resulted consistent with the way they are formed in the years 1987-2006.

are represented by motoring budget shares that increased from 9% to 14% and leisure services that increased from 4% to 11%, a sign of an increased level of welfare in English population.

In table 2, we consider the same averages but for more homogeneous classes of normalized total expenditure x_h (as a proxy for income): poor ($x_h \leq 30$), medium ($30 < x_h \leq 70$), and rich ($x_h > 70$) households. While the same time patterns highlighted above remain true for all classes, we find differences among households with different income level. In each 10-years window considered poor households allocate more budget than rich to necessities as food (25% against 14% in the last window) and fuel, light, and power (6% against 2% in the last window, and up to 11% against 5% in the first window): this is the well known Engel's law. Also poor households allocate less budget than rich to motoring (10% against 17% in the last window) and leisure services (9% against 14% in the last window). Finally, irrespectively of their income households allocate between 15% and 17% of their budget to housing and between 4% and 5% to alcoholic drinks.

Already from this descriptive analysis we can classify goods according to their budget shares into three broad classes: necessities (budget shares decreasing with total expenditure), luxuries (budget shares increasing with total expenditure), and goods for which the budget share is constant with respect to total expenditure.

5 Results

5.1 Number of factors

Table 3 displays the estimates of the number of factors for different time windows. We consider time windows of 3, 5, and 10 years length. Due to the asymptotic properties of the criteria employed we cannot consider the results concerning the 3 years windows reliable enough, since in this case $J = 39$, which is definitely too small. Results concerning 5 and 10 years windows ($J = 65$ and $J = 130$) are more reliable and indeed are more homogeneous. We find in this case between 3 and 2 common factors. We thus carry on the identification analysis that follows by considering the maximum number of factors allowed by the criteria employed, i.e. $R = 3$.

In the last 3 columns of table 3 we show the proportion of variance explained by each factor. The first factor explains, for all the time windows considered, always more than 50% of total variance, being clearly the most important. Its contribution to the total variance of budget shares, however, decreases with time. This is probably due to the fact that in the last thirty or forty years families of the same income class have increasingly differentiated their consumption habits, so that idiosyncratic components have played a relatively bigger role. This may in turn be due to a wider range of products available joint with an increase in families total resources. Moreover, as the first factor will be interpreted as related to necessities (see section 5.2), a decrease in the explained variance of the first factor can also be seen as a sign of an increased level of welfare, as mentioned above.

Factor models are identified under a specific condition on diverging eigenvalues of the covariance matrix of the data (see Assumption 3). This is precisely the Assumption tested by the Bai and Ng (2002) criterion which shows evidence of an additional one or even two less important, but still common, factors explaining a much lower proportion of variance, in fact lower than 10%. We must stress the fact that not recognizing the existence of such factors would imply the existence of common features in the idiosyncratic components. Indeed, in order to be truly common the factors do not have to be necessarily large (a relative concept) in terms of explained variance, but they have to be pervasive, a well defined feature that can be measured by studying the asymptotic behaviour of eigenvalues. This is exactly what the employed criteria do.

5.2 Interpretation of the factors

In this section we present results only for the last window considered, i.e. from 1997 to 2006.[6] The identification of the factors is based on the independent component analysis, as explained in section 3.4. This method can be applied only if the underlying independent components, and, consequently, the estimated (non-identified) factors are non-Gaussian. Figure 1 shows the quantiles of estimated factors *vs.* Gaussian quantiles: a non-linear relation clearly appears. This suggests that the factors do not follow a Gaussian distribution. We also test directly for Gaussianity. The Shapiro-Wilk test rejects the hypothesis of Gaussianity at the 0.05 level of significance in each case. The resulting p-values are: 6.9 10^{-6}, 2.6 10^{-9}, 1.5 10^{-3} for the three factors estimated via PCA and 8.4 10^{-9}, 2.3 10^{-5}, 1.6 10^{-11} for the identified factors.[7]

Once estimation and identification of the factors are completed, we obtain the model:

$$w_{jh} = \sum_{r=1}^{3} a_{jr} \widetilde{f}_{rh} = \sum_{r=1}^{3} a_{jr} g_r(x_h), \quad j = 1, \ldots, J \quad h = 1, \ldots, H, \tag{9}$$

where g_r are three non-linear functions of total expenditure x_h, which are still unknown.

In order to estimate g_r we regress the estimated factor $\widetilde{f}_{rh}$ on total expenditure, for each $r = 1, 2, 3$:

$$\widetilde{f}_{rh} = g_r(x_h) + \varepsilon_{rh}, \quad h = 1, \ldots H. \tag{10}$$

Figure 2 displays the three factors $\widetilde{f}_{rh}$ as functions of total expenditure together with their estimated non-parametric fits $\widehat{g}_r(x_h)$, obtained by means of the Nadaraya-Watson kernel regression. The first function $\widehat{g}_1(x_h)$ decreases for small values of total expenditure and then remains stable. This pattern is very similar to the pattern of food and fuel budget shares, as evidenced from figure 3 (a-b). Table 4 displays the estimates Pearson correlation coefficients between the three factors and budget shares.[8] As expected, we find that the first factor is highly correlated with food and fuel budget shares (correlation coefficient 0.84 and 0.80 respectively). This again suggests that the first factor captures consumption

[6] Results for the other windows considered are available upon request.

[7] The Shapiro-Francia test produces analogous results.

[8] Spearman and Kendall rank correlation coefficients produce analogue results.

patterns typically associated with the Engel's law: as income (proxied by total expenditure) rises, budget shares decrease, the falling down of budget share being more dramatic for low levels of income.

The second function, $\widehat{g}_2(x_h)$, as shown in figure 2 (b), apart from the very first portion of total expenditure, is increasing with total expenditure. It is associated with categories of expenditure that are more likely to include luxuries as clothing and footwear, motoring, and leisure services. Indeed, from figure 3 (c-d) we see that the second factor displays a pattern similar to leisure service and motoring budget shares. These are also the budget shares with which the second factor is mostly correlated (see table 4).

Finally, the third function, $\widehat{g}_3(x_h)$, is slightly increasing in the first quarter of total expenditure and then slightly decreasing, remaining on average approximately constant. This pattern is similar to the one displayed by housing (see figure 3 e), which is the budget share with which the third factor is most correlated (see table 4). The third factor tends to reflect patterns of intermediate categories, that is goods that have quite stable budget shares over total expenditure.

In order to compare our results with the literature (Lewbel, 1991; Banks et al., 1997), we also investigate which functional form of total expenditure better fits each identified factor. We estimate the following functions of total expenditure x_h: x_h, x_h^2, x_h^{-1}, x_h^{-2}, $\log x_h$, $(\log x_h)^2$, $x_h \log x_h$. These are are the functional forms also considered by Lewbel (1991, p. 719) and Donald (1997, pp. 122-123). As displayed in table 5, the first factor obtains the best fit, in terms of R^2, with the simple logarithmic form: $\alpha + \beta \log x$. This is the functional form incorporated in the Working-Leser model. The second and third factors (see tables 6 and 7) obtain the best relative fit, in terms of R^2, with the quadratic form $\alpha + \beta x^2$. Notice, however, that, as regards the third factor, R^2 are quite small for all the functional forms (many of which result to be non-significant), so that a constant relation constitutes a good approximation. This is also confirmed from the analysis of the third factor when excluding the highest 20% of high income family. The quadratic fit becomes a constant.

In sum the parametric specification of the system of Engel curves which is most consistent with our findings is:

$$w_{jh} = a_j + b_j \log x_h + c_j x_h^2 + e_{jh}, \quad j = 1, \ldots, J, \quad h = 1, \ldots, H. \tag{11}$$

This is consistent with Lewbel (1997), who proposed:

$$w_{jh} = a_j + b_j \log x_h + c_j \phi(x_h) + e_{jh}, \quad j = 1, \ldots, J, \quad h = 1, \ldots, H, \tag{12}$$

for some non-linear function ϕ. Banks et al. (1997), using 1980-1982 U.K. FES data, found that Engel curves have indeed the form of equation (12), with $\phi(x_h) = (\log x_h)^2$. In this latter respect, our results slightly differ from previous findings, since our last term is quadratic in x_h.

A final way to interpret the factors is based on the estimation of the average derivative of $g_r(x_h)$, whose sign is strictly connected to whether a category of expenditure should be classified as luxury or necessity. Total expenditure elasticity has a direct connection with

the double log model, since for any good j:

$$\log w_{jh} = \alpha_{jh} + (\epsilon_j - 1)\log x_h, \quad j = 1, \ldots, J,\ h = 1, \ldots, H, \tag{13}$$

where ϵ_j is the the total expenditure elasticity of good j (see Deaton and Muellbauer, 1980, p. 17). Hence, if a factor $\widetilde{f}_{rh}$ is supposed to represent a necessity, we should expect that the average derivative of the factor with respect to $\log x_h$ is less than one (and greater than one if it represents a luxury). After rescaling the factor in such a way that $\widetilde{f}_{rh} > 0$, we estimate the average (over households) derivative $\frac{\partial \widetilde{f}_{rh}}{\partial x_h}$, since it has the same sign as $\frac{\partial \log \widetilde{f}_{rh}}{\partial \log x_h}$, being in this case both $\widetilde{f}_{rh}$ and x_h greater than zero.

We estimate average derivatives in a non-parametric manner, using the method proposed by Härdle and Stoker (1989), which being based on kernel density estimates does not presuppose any functional form of the factors. Table 8 displays the estimated average derivatives together with results from the Wald test for zero derivative. The null hypothesis is rejected at a 5% level of significance for the first and second factors. The signs and the significance of the derivatives confirm that the first factor captures necessities, the second factor captures luxuries, while the third factor captures goods with income elasticity close to unit, i.e. zero derivative.

6 Conclusions

In this paper, we propose a method to determine the rank of a system of Engel curves for different categories of expenditures expressed in budget shares form. The rank of such a system determines the maximum number of functions of total expenditure that drive consumers' behavior. The method we propose is based on approximate factor models and independent component analysis. We frame the problem of finding the rank as the problem of determining the number of latent common factors that explain variations of the system of budget shares. Herein, we identify the maximum number of common factors by means of the criteria proposed by Bai and Ng (2002). The factors can be estimated via approximate principal components and then identified by independent component analysis.

We apply this method to U.K. Family Expenditure Survey annual data. In order to apply factor analysis, we build a large dimension panel of data, in which the budget shares (relative to 13 categories of expenditures) of 100 representative households are pooled over different years. The way this data set is built is based on the method to pool and normalize expenditures over years proposed by Kneip (1994). This large dimensional dataset permits us to eschew any assumption of uncorrelation among idiosyncratic shocks. The departure from normal distribution that budget shares display and a hypothesis about the nature of the fundamental drivers of consumption decisions permit us to apply independent component analysis to achieve identification.

Once the common latent factors are identified, we study their properties by different parametric and non-parametric regressions. We also estimate their average derivative by applying the method proposed by Härdle and Stoker (1989). Results show that the

system of Engel curves should be specified as the sum of a logarithmic, quadratic, and constant term, in a form which is consistent with the model suggested by Lewbel (1997). Moreover, the three common factors reflect consumption behaviors which are typical of necessities, luxuries, and unity elasticity goods.

References

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate static factor models. *Statistics and Probability Letters*, 80.

Aversi, R., Dosi, G., Fagiolo, G., Meacci, M., and Olivetti, C. (1999). Demand dynamics with socially evolving preferences. *Industrial and Corporate Change*, 8:353–468.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221.

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

Belouchrani, A., Abed Meraim, K., Cardoso, J., and Moulines, E. (1997). A blind source separation tecnique based on second order statistics. *IEEE Transactions on Signal Processing*, 45.

Bonhomme, S. and Robin, J. (2009). Consistent noisy independent component analysis. *Journal of Econometrics*, 149:12–25.

Cardoso, J. and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings part F Radar and Signal Processing*, 140:362–362.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51:1305–1324.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36:287–314.

Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, 70:312–326.

Donald, S. (1997). Inference concerning the number of factors in a multivariate nonparamentric relationship. *Econometrica*, 65:103–132.

Fagiolo, G., Alessi, L., Barigozzi, M., and Capasso, M. (2010). On the distributional properties of household consumption expenditures: The case of Italy. *Empirical Economics*, 38.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82:540–554.

Gorman, W. M. (1981). Some Engel curves. In Deaton, A., editor, *Essays in the Theory and Measurements of Consumer Behaviour in Honor of Sir Richard Stone*. Cambridge University Press.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84:986–995.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *The Annals of Statistics*, 22:1386–1427.

Lewbel, A. (1991). The rank of demand systems: Theory and nonparametric estimation. *Econometrica*, 59:711–730.

Lewbel, A. (1997). Consumer demand systems and household equivalence scales. *Handbook of Applied Econometrics: Microeconomics*, 2:167–201.

Onatski, A. (2010). Determining the number of factors form empirical distribution of eigenvalues. *Review of Economics and Statistics*. forthcoming.

Stock, J. and Watson, M. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, pages 351–394.

Tables and figures

Table 1: Average budget shares over all household income classes.

Good	**window**		
	1977-1986	**1987-1996**	**1997-2006**
Housing	0.17	0.16	0.16
Fuel, light, power	0.07	0.06	0.04
Food	0.24	0.21	0.19
Alcoholic drinks	0.05	0.05	0.04
Tobacco	0.04	0.03	0.02
Clothing and footwear	0.07	0.06	0.05
Household goods	0.06	0.08	0.08
Household services	0.03	0.05	0.05
Personal goods and services	0.03	0.04	0.04
Motoring	0.09	0.12	0.14
Fares and other travel	0.02	0.02	0.02
Leisure goods	0.04	0.05	0.05
Leisure services	0.04	0.08	0.11

Table 2: Average budget shares for different household income classes.

Good	**window**									
	1977-1986			**1987-1996**			**1997-2006**			
	poor	**medium**	**rich**	**poor**	**medium**	**rich**	**poor**	**medium**		**rich**
Housing	0.19	0.16	0.14	0.17	0.17	0.15	0.17	0.17		0.15
Fuel, light, power	0.11	0.06	0.05	0.09	0.05	0.04	0.06	0.03		0.02
Food	0.32	0.23	0.18	0.28	0.20	0.16	0.25	0.17		0.14
Alcoholic drink	0.04	0.06	0.05	0.04	0.05	0.05	0.04	0.04		0.04
Tobacco	0.05	0.03	0.02	0.04	0.02	0.02	0.03	0.02		0.01
Clothing and footwear	0.04	0.07	0.09	0.04	0.06	0.07	0.05	0.06		0.06
Household goods	0.04	0.06	0.08	0.07	0.08	0.09	0.07	0.08		0.09
Household services	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.05		0.05
Personal goods and services	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04		0.04
Motoring	0.04	0.10	0.12	0.07	0.13	0.15	0.10	0.15		0.17
Fares and other travel	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02		0.02
Leisure goods	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.05		0.05
Leisure services	0.02	0.04	0.05	0.05	0.08	0.10	0.09	0.11		0.14

poor: households with normalized total expenditure $x_h \leq 30$; **medium**: households with normalized total expenditure $30 < x_h \leq 70$; **rich**: households with normalized total expenditure $x_h > 70$.

Table 3: Determining the number of factors and their explained variance.

window	IC_1	IC_2	ABC	O	Avg.	Explained Variance		
						Factor 1	Factor 2	Factor 3
1968 - 1970	3	2	2	1	2.00	0.74	0.04	0.03
1971 - 1973	3	3	1	1	2.00	0.73	0.04	0.03
1974 - 1976	6	3	2	2	3.25	0.69	0.08	0.03
1977 - 1979	12	7	1	1	5.25	0.70	0.05	0.04
1980 - 1982	5	4	1	1	2.75	0.70	0.04	0.04
1983 - 1985	10	5	1	1	4.25	0.74	0.05	0.04
1986 - 1988	12	2	1	2	4.25	0.62	0.09	0.03
1989 - 1991	15	7	1	1	6.00	0.62	0.06	0.05
1992 - 1994	15	8	1	1	6.25	0.59	0.06	0.05
1995 - 1997	8	6	1	1	4.00	0.60	0.05	0.04
1998 - 2000	13	5	2	1	5.25	0.58	0.06	0.05
2001 - 2003	13	7	1	1	5.50	0.60	0.05	0.04
2004 - 2006	13	13	2	1	7.25	0.53	0.07	0.06
1972 - 1976	3	3	3	3	3.00	0.70	0.05	0.03
1977 - 1981	6	2	1	1	2.50	0.69	0.03	0.03
1982 - 1986	4	4	1	2	2.75	0.71	0.05	0.03
1987 - 1991	9	4	3	2	4.50	0.62	0.06	0.04
1992 - 1996	6	2	3	1	3.00	0.59	0.04	0.04
1997 - 2001	5	4	2	1	3.00	0.58	0.04	0.04
2002 - 2006	8	6	1	1	4.00	0.55	0.05	0.04
1977 - 1986	5	2	2	2	2.75	0.70	0.03	0.02
1987 - 1996	7	2	2	2	3.25	0.60	0.04	0.03
1997 - 2006	3	2	4	2	2.75	0.56	0.04	0.03

$\mathbf{IC_1}$ and $\mathbf{IC_2}$: Bai and Ng (2002) criteria. **ABC**: Alessi et al. (2010) criterion. **O**: Onatski (2010) criterion. **Avg.**: average computed over the four criteria. **Explained variance**: variance explained by each factor computed with respect to total variance.

Table 4: Correlations between factors and budget shares: 1997-2006.

Good	**Average Correlation**		
	Factor 1	**Factor 2**	**Factor 3**
Housing	-0.12	-0.41	0.27
Fuel, light, power	0.84	-0.43	0.14
Food	0.80	-0.50	0.20
Alcoholic drinks	-0.31	-0.01	0.08
Tobacco	0.65	-0.39	0.14
Clothing and footwear	-0.29	0.31	-0.07
Household goods	-0.21	0.23	-0.16
Household services	0.10	0.02	-0.01
Personal goods and services	-0.16	0.15	-0.08
Motoring	-0.66	0.42	-0.11
Fares and other travel	-0.04	0.22	-0.08
Leisure goods	-0.10	0.10	-0.11
Leisure services	-0.46	0.41	-0.25

Averages are computed over the 10 years period 1997-2006.

Table 5: Parametric fits for the first factor: 1997-2006.

	Functional form		α	β	**adj.R^2**
Factor 1	*non-parametric*				0.86
	$\alpha + \beta x_h$	coeff	***	***	0.42
		t-stat	7.35	-8.47	
	$\alpha + \beta x_h^2$	coeff	***	***	0.22
		t-stat	4.05	-5.42	
	$\alpha + \beta x_h^{-1}$	coeff	***	***	0.51
		t-stat	-4.16	10.25	
	$\alpha + \beta x_h^{-2}$	coeff	$\circ$	***	0.22
		t-stat	-0.87	5.51	
	$\alpha + \beta \log(x_h)$	coeff	***	***	0.74
		t-stat	16.37	-16.78	
	$\alpha + \beta(\log(x_h))^2$	coeff	***	***	0.63
		t-stat	12.24	-14.40	
	$\alpha + \beta x(\log(x_h))$	coeff	***	***	0.36
		t-stat	6.25	-7.51	

Regressions of the factors on functions of total expenditure x_h for $h = 1, \ldots, H$. Symbols report whether coefficients are significant at the *** 0.01, ** 0.05, * 0.1 level ($\circ$ = no significance at any level < 0.01)

Table 6: Parametric fits for the second factor: 1997-2006.

	Functional form		α	β	**adj.R^2**
Factor 2	*non-parametric*				0.57
	$\alpha + \beta x_h$	signif	***	***	0.44
		t-stat	-7.64	8.80	
	$\alpha + \beta x_h^2$	signif	***	***	0.50
		t-stat	-7.52	10.06	
	$\alpha + \beta x_h^{-1}$	signif	∘	∘	0.00
		t-stat	0.31	-0.75	
	$\alpha + \beta x_h^{-2}$	signif	∘	∘	0.00
		t-stat	-0.05	0.29	
	$\alpha + \beta \log(x_h)$	signif	***	***	0.20
		t-stat	-4.76	4.91	
	$\alpha + \beta(\log(x_h))^2$	signif	***	***	0.30
		t-stat	-6.14	7.08	
	$\alpha + \beta x(\log(x_h))$	signif	***	***	0.46
		t-stat	-7.73	9.30	

Regressions of the factors on functions of total expenditure x_h for $h = 1, \ldots, H$. Symbols report whether coefficients are significant at the *** 0.01, ** 0.05, * 0.1 level
(∘ = no significance at any level < 0.01)

Table 7: Parametric fits for the third factor: 1997-2006.

	Functional form		α	β	**adj.R**^{2}
Factor 3	*non-parametric*				0.23
	$\alpha + \beta x_h$	signif	***	***	0.09
		t-stat	2.83	-3.26	
	$\alpha + \beta x_h^2$	signif	***	***	0.14
		t-stat	3.09	-4.14	
	$\alpha + \beta x_h^{-1}$	signif	∘	∘	0.00
		t-stat	0.12	-0.30	
	$\alpha + \beta x_h^{-2}$	signif	∘	∘	0.00
		t-stat	0.08	-0.48	
	$\alpha + \beta \log(x_h)$	signif	∘	∘	0.01
		t-stat	1.39	-1.43	
	$\alpha + \beta(\log(x_h))^2$	signif	**	**	0.04
		t-stat	2.02	-2.23	
	$\alpha + \beta x(\log(x_h))$	signif	***	***	0.10
		t-stat	2.93	-3.53	

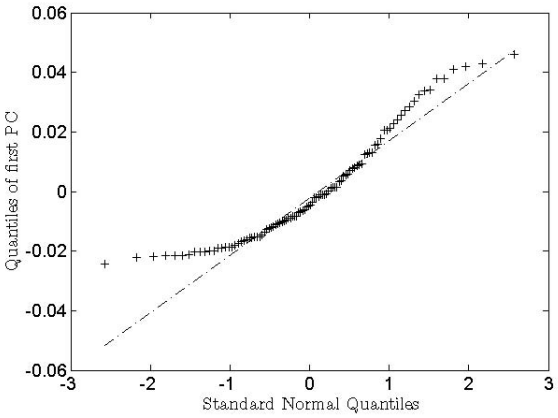
Regressions of the factors on functions of total expenditure x_h for $h = 1, \ldots, H$. Symbols report whether coefficients are significant at the *** 0.01, ** 0.05, * 0.1 level (∘ = no significance at any level < 0.01)

Table 8: Average derivatives: 1997-2006.

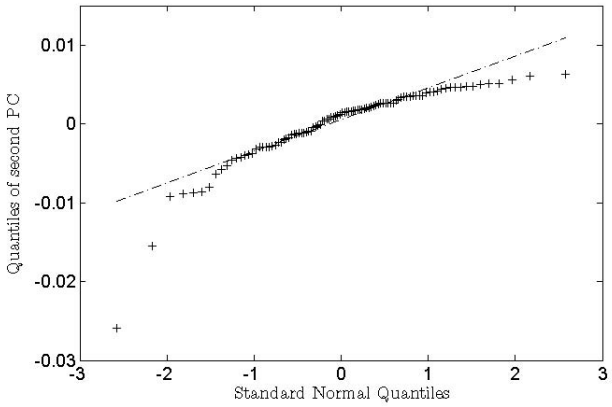
Factor 1	-0.3159*
standard error	0.1450
Wald statistic	4.7430
Factor 2	0.2830**
standard error	0.0966
Wald statistic	8.5802
Factor 3	-0.0783
standard error	0.1082
Wald statistic	0.5229

Average derivative obtained with the Härdle-Stoker (1989) method. Standard errors are obtained via bootstrap procedure. Wald Test with H_0 : average derivative = 0. Significance at the * 0.05, ** 0.01 level.

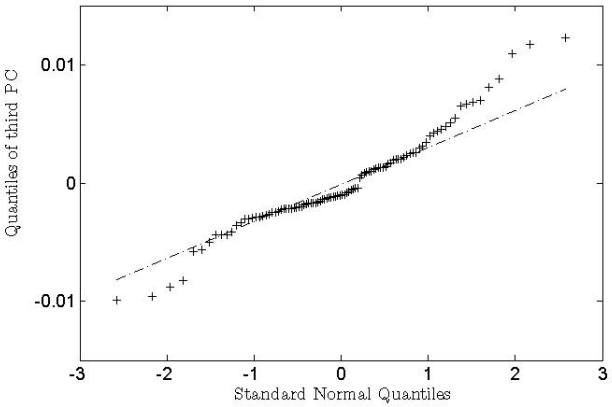
Figure 1: Q-Q plots of the three estimated factors for the window 1997-2006.



(a) First factor

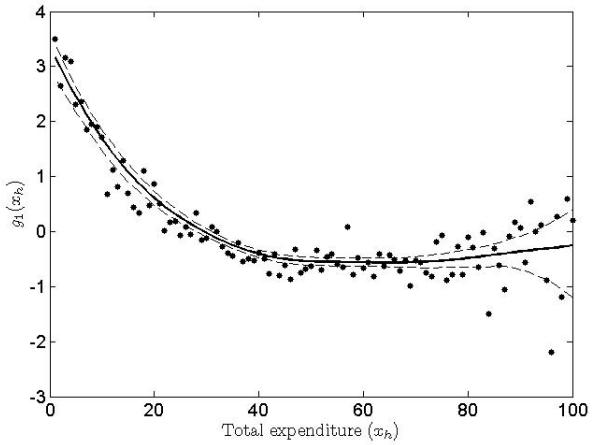


(b) Second factor

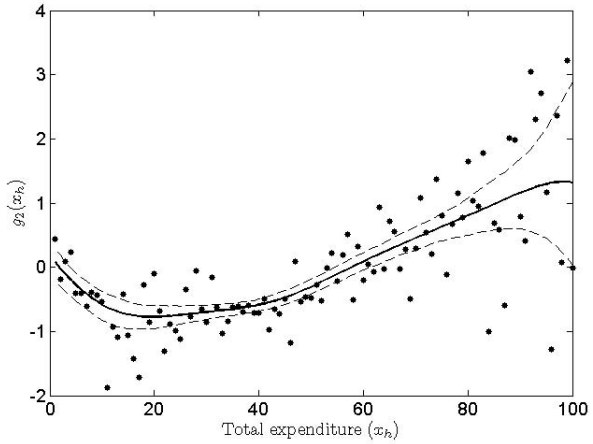


(c) Third factor

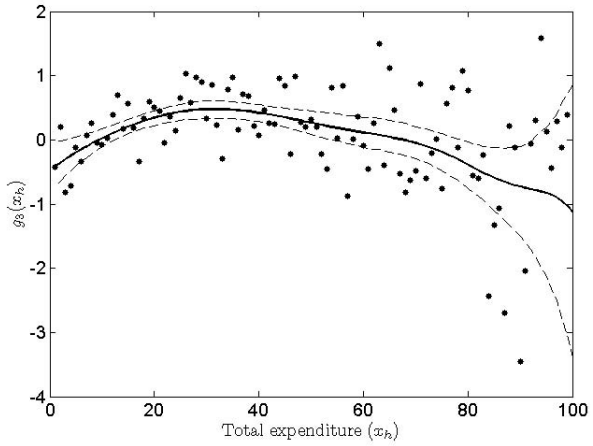
Figure 2: Estimated non-parametric fits: 1997-2006.



(a) First factor



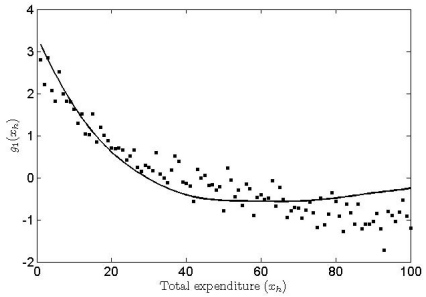
(b) Second factor



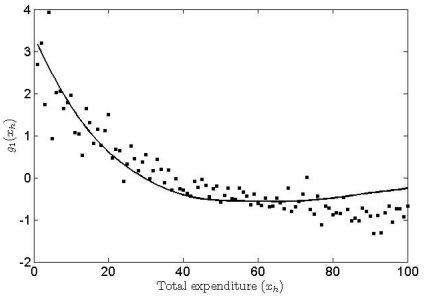
(c) Third factor

Non-parametric estimates (Nadaraya-Watson regressions) of the functions of total expenditure (solid line): $g_r(x_h)$, and 95% confidence intervals (dashed lines). Points denote the values taken by the factors. Notice that by construction factors have mean zero.

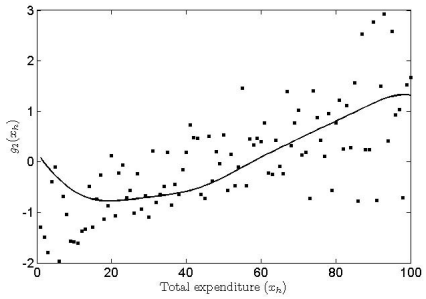
Figure 3: Interpreting the factors: 1997-2006.



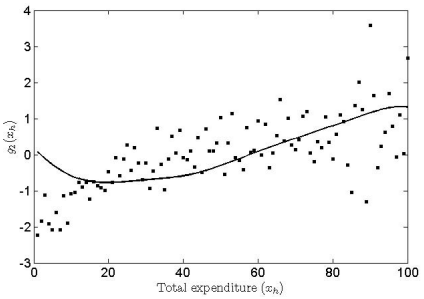
(a) First factor and food BS



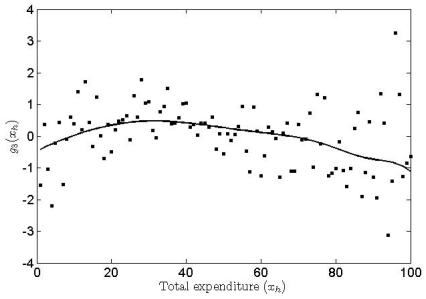
(b) First factor and fuel, light, power BS



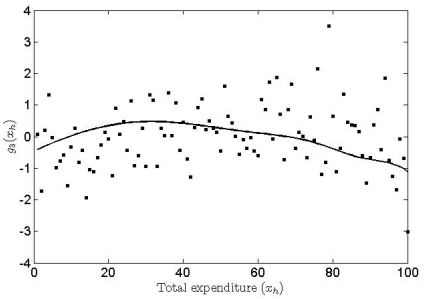
(c) Second factor and leisure services BS



(d) Second factor and motoring BS



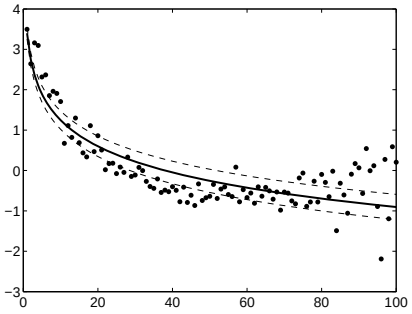
(e) Third factor and housing BS



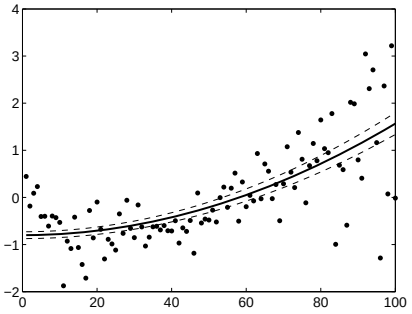
(f) Third factor and alcoholic drinks BS

Scatter plots of budget shares w_{jh} of selected goods (squares) and estimated non-parametric functions of total expenditure $g_r(x_h)$ (solid lines), as obtained from Nadaraya-Watson regressions of the estimated factors on x_h. Notice that by construction the non-parametric fits have mean zero, hence w_{jh} are rescaled accordingly.

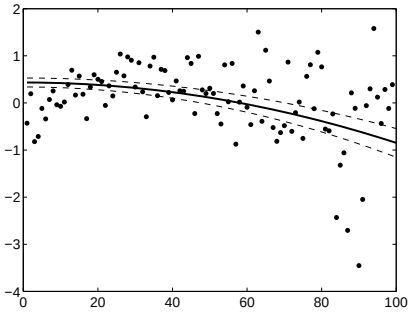
Figure 4: Estimated parametric fits: 1997-2006.



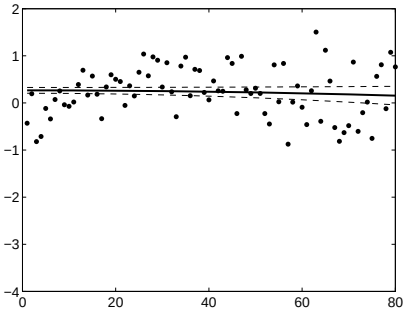
(a) First factor - fit



(b) Second factor - fit



(c) Third factor - fit



(d) Third factor - fit without rich households

Parametric estimates (non-linear regressions) of the functions of total expenditure (solid line): (a) $g_1(x_h) = \alpha + \beta \log(x_h)$, (b) $g_2(x_h) = \alpha + \beta(x_h)^2$, (c) $g_3(x_h) = \alpha + \beta(x_h)^2$, (d) $g_3(x_h) = \alpha + \beta(x_h)^2$ without 20% richest households $x_h = 1, \ldots 80$, and 95% confidence intervals (dashed lines). Notice that by construction the factors and the non-linear fits have mean zero.