KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and
Applied Economics

Department of Economics

Nonparametric Analysis of Household Labour Supply: Goodness-of-Fit and Power of the Unitary and the Collective Model

by

Laurens CHERCHYE
Frederic VERMEULEN

Public Economics

Center for Economic Studies
Discussions Paper Series (DPS) 03.02
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

January 2003

DISCUSSION
PAPER

Nonparametric analysis of household labour supply:
Goodness-of-fit and power of the unitary and the collective model[*]

Laurens Cherchye[†] and Frederic Vermeulen[‡]

January 2003

Abstract

We compare the empirical performance of the unitary and the collective approach to modelling observed labour supply behaviour. Deviating from the mainstream literature, we conduct a nonparametric analysis, which avoids the distortive impact of an erroneously specified functional form for the preferences and/or the intrahousehold bargaining process. Our analysis specifically focuses on the goodness-of-fit of the two behavioural models. To guarantee a fair comparison, we complement this goodness-of-fit analysis with a power analysis. Our results strongly favour the collective approach to modelling the behaviour of multi-person households. More generally, they illustrate the usefulness of nonparametric testing tools for the empirical evaluation of theoretical behavioural models.

Key words: labour supply, collective model, unitary model, nonparametric tests, revealed preferences.

JEL-classification: C14, D12, J22.

[*] We are grateful to Denis Beninger, Martin Browning and Erik Schokkaert for useful comments.

[†] University of Leuven Campus Kortrijk and Center for Economic Studies, E. Sabbelaan 53, B-8500 Kortrijk, Belgium, e-mail: laurens.cherchye@kulak.ac.be. Postdoctoral Fellow of the Fund for Scientific Research - Flanders.

[‡] Center for Economic Studies, University of Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, e-mail: frederic.vermeulen@econ.kuleuven.ac.be.

1 Introduction

Standard microeconomic theory assumes that a household acts as if it were a single decision maker. Within this tradition, household demand is assumed to result from maximizing a unique utility function subject to a household budget constraint. However, a growing body of evidence suggests that this standard *unitary* model is at odds with observed household behaviour; the associated restrictions of homogeneity, symmetry and negativity have been rejected at numerous occasions (e.g., Blundell, 1988).

A more recent alternative, the so-called *collective* approach to household behaviour (Chiappori, 1988, 1992), explicitly takes account of the fact that multi-person households consist of several individuals with their own rational preferences; household decisions are then the Pareto efficient outcomes of a bargaining process. This collective approach entails other behavioural restrictions than the unitary model. Interestingly enough, these restrictions seem to better fit the data than the 'traditional' unitary restrictions; e.g., Browning *et alii* (1994), Fortin and Lacroix (1997), Browning and Chiappori (1998) and Chiappori *et alii* (2002).

Still, the hitherto employed tests of the unitary and collective models are *parametric* in nature. Hence, they crucially depend on the functional form that is used for representing the preferences and/or the intrahousehold bargaining process. They do not only test the unitary or collective approach as such, but also an *ad hoc* functional specification; rejecting the unitary restrictions may well be due to ill-specification.

Nonparametric tests for consistency of observed behaviour with utility maximization or Pareto efficiency do not require any assumptions regarding the parametric form of utility functions or the intrahousehold bargaining process; see, e.g., Afriat (1967), Varian (1982), Chiappori (1988) and Snyder (2000). These tests are directly based on revealed preference theory, which makes them particularly attractive for testing consistency of the data with theoretical behavioural models.

This suggests using nonparametric testing tools for comparing the empirical performance of the unitary and collective model for household behaviour; the validity of the findings would no longer be 'obscured' by non-verifiable parametric assumptions. However, to the best of our knowledge -and in fact somewhat surprisingly-, an in-depth nonparametric comparison has not yet been carried out. This paper wants to fill that gap in the existing literature, by studying the specific case of household labour supply behaviour. Conveniently, our focus on labour supply also guarantees substantial price/wage variation across individuals, which can only benefit the empirical comparison.

Our following assessment specifically concentrates on two types of empirical performance measures:
- First, we compute nonparametric *goodness-of-fit* measures associated with each behavioural model. Varian (1990, 1993; based on Afriat, 1972; see also Cox, 1997) has developed such nonparametric measures for 'the degree to which empirical behaviour is consistent with theoretical behaviour'. (These goodness-

of-fit measures also account for possible reporting errors in the data, which is an important practical consideration in empirical applications.) Goodness-of-fit measures correct for the 'sharp' nature of standard nonparametric tests, which only tell us whether or not a given dataset is *exactly* consistent with the behavioural model subject to testing. In our opinion, these measures are directly applicable for assessing and comparing the empirical performance of theoretical behavioural models.

- Second, we calculate measures for the *power* of the consistency tests for each model, i.e. the probability of detecting the alternative hypothesis (e.g., based on Becker's (1962) notion of irrational behaviour); Bronars (1987; also Cox, 1997) first proposed nonparametric power measures. We believe that a full (and fair) comparison of the two behavioural models under study should complement the above goodness-of-fit analysis with a power analysis. (A similar point was, e.g., raised by Snyder (2000).) This is all the more a valid concern since a frequently cited weakness of nonparametric tools is that they have low power.[1] A careful power assessment allows us to counter that criticism. (In fact, this so-called lack of power may at least partly explain why nonparametric consistency tests are so rarely employed in practice, despite their attractive feature of imposing minimal structure.)

Our empirical evaluation uses a cross-section dataset of Belgian households (consisting of working individuals), which we divide in three subsamples: female singles, male singles and couples. We essentially discuss two types of comparisons:

- First, we compare the empirical performance of the unitary model for singles with that for couples. The rationale of this comparison is that the standard unitary approach should always be fully applicable to singles, even if it does not well fit the observed behaviour of couples; for singles the behavioural implications of the unitary model coincide with those of the collective model. This first comparison should give us a deeper understanding of the harmless/harmful nature of the aggregation assumptions that underlie the unitary modelling of couples' behaviour.
- Second, we compare the empirical results of the collective model with those of the unitary model, both applied to the data of couples. Because the collective and unitary models evidently have different implications for couples' behaviour, these results should give us better insight into which of the models does the better job in explaining the data.

The remainder of the paper unfolds as follows. Section 2 briefly reviews the nonparametric methodology for testing the unitary and the collective labour supply models. In addition, we introduce the nonparametric goodness-of-fit and power measures. Section 3 presents the results of our application to Belgian household data. Section 4 concludes.

[1] Blundell *et alii* (2001) propose methodological extensions that can considerably improve the power of the nonparametric tests. Their proposal crucially builds on groups of different households facing the same relative prices. However, a typical feature of our (labour supply) setting is precisely that prices/wages substantially vary over households, which makes the Blundell *et alii* approach not directly applicable in the current context.

2 Methodology

2.1 Testing the unitary model

For the sake of compactness, we only discuss unitary consistency tests for couples with two working individuals (A and B). Our discussion is directly translated to the singles' case.

The nonparametric approach starts from n observations for household consumption and the household members' labour supply. For each household i ($i = 1, ..., n$) we denote the net wage rate and leisure amount of individual I ($I = A, B$) by w_i^I and l_i^I, respectively. (The leisure amount is computed from observed labour supply $\ell_i^I = T - l_i^I$, with T the individuals' time endowment.) Next, we use y_i and c_i to respectively denote household i's (total) nonlabour income and (total) consumption. Finally, we represent the set of all observations by $S = \left\{\left(c_i, l_i^A, l_i^B, w_i^A, w_i^B, y_i\right), i = 1, ..., n\right\}$.

Within the unitary model, the household decision problem boils down to maximizing (at the household level) a nonsatiated utility function $v\left(c_i, l_i^A, l_i^B\right)$ subject to the household budget constraint $c_i + w_i^A l_i^A + w_i^B l_i^B \leq y_i + w_i^A T + w_i^B T$; without losing generality, we set the price of consumption to 1. A necessary and sufficient condition for the data to be consistent with this utility maximization problem is that there exists a utility function that *rationalizes* the household data, i.e.:

Definition 1 *A utility function v* rationalizes *the observed set S if for all $i \in \{1, ..., n\}$: $v\left(c_i, l_i^A, l_i^B\right) \geq v\left(c, l^A, l^B\right)$ for all $\left(c, l^A, l^B\right)$ such that $c_i + w_i^A l_i^A + w_i^B l_i^B \geq c + w_i^A l^A + w_i^B l^B$.*

Varian (1982) has demonstrated that there exists such a data rationalizing utility function if and only if the observed set S is consistent with the *generalized axiom of revealed preference* (*GARP*). To formally state this last consistency condition, we first need the following revealed preference definition (using $\left(1, w^A, w^B\right)' = \mathbf{w}$ and $\left(c, l^A, l^B\right)' = \mathbf{l}$):

Definition 2 *An observation $\mathbf{l}_i$ is* revealed preferred *to a bundle $\mathbf{l}$, denoted by $\mathbf{l}_i R \mathbf{l}$, if $\mathbf{w}_i' \mathbf{l}_i \geq \mathbf{w}_i' \mathbf{l}_j$, $\mathbf{w}_j' \mathbf{l}_j \geq \mathbf{w}_j' \mathbf{l}_k$, ..., $\mathbf{w}_m' \mathbf{l}_m \geq \mathbf{w}_m' \mathbf{l}$ for some sequence of observations $(\mathbf{l}_i, \mathbf{l}_j, ..., \mathbf{l}_m)$.*

We can now define the GARP condition as:

Definition 3 *The observed set S satisfies GARP if for all $i, j \in \{1, ..., n\}$: $\mathbf{l}_i R \mathbf{l}_j$ then $\mathbf{w}_j' \mathbf{l}_j \leq \mathbf{w}_j' \mathbf{l}_i$.*

To facilitate our further exposition, we rephrase this definition as follows:

Definition 4 *The observed set S satisfies GARP if for all $j \in \{1, ..., n\}$: $\mathbf{w}_j' \mathbf{l}_j = \min_{\mathbf{l} \in \mathbf{RP}_j} \mathbf{w}_j' \mathbf{l}$ for $RP_j = \{\mathbf{l}_i : \mathbf{l}_i R \mathbf{l}_j; i \in \{1, ..., n\}\}$.*

This last version directly expresses the idea that observation $j \in \{1, ..., n\}$ is (theoretically) utility maximizing under its budget constraint if and only if it is expenditure minimizing over its 'better than' set; in the (empirical) GARP condition this last set is approximated by the 'revealed preferred' set RP_j.

Consistency of S with GARP (and thus with a rationalizing utility function) is easily tested: we first identify the set RP_j and consequently check the expenditure minimization condition. See, e.g., Varian (1982; p. 949) for an efficient algorithm.

2.2 Testing the collective model

The collective approach essentially differs from the unitary approach in that each household member is characterized by own rational preferences, with household decisions resulting from a Pareto efficient bargaining process (Chiappori, 1988, 1992). Although the individuals' preferences can be very general, we restrict attention to *egoistic* preferences in our discussion; preferences only depend on own (private) consumption and leisure.[2] Empirically, the modelling of this collective approach is somewhat more involved as the private consumption of each household member is usually not observed; labour supply datasets only reveal information on total household consumption.

To see how the empirical analysis of household behaviour proceeds (even under limited information), we first introduce some additional notation. We denote individual I's private consumption by c_i^I, and the vectors $\left(1, w_i^I\right)'$ and $\left(c_i^I, l_i^I\right)'$ by respectively $\mathbf{w}_i^I$ and $\mathbf{l}_i^I$ $(I = A, B)$.

Now consider the case where a (two-person) household is characterized by a pair of (nonsatiated) utility functions, $v^A\left(c_i^A, l_i^A\right)$ and $v^B\left(c_i^B, l_i^B\right)$, *and* a sharing rule $\phi\left(w_i^A, w_i^B, y_i\right)$ which determines the distribution of the household's nonlabour income y_i over the household members.[3] This sharing rule is formally defined as follows (see, e.g., Chiappori, 1988, 1992).

Definition 5 *A* sharing rule ϕ *is a function which maps the vector* $\left(w_i^A, w_i^B, y_i\right)'$ *to* $\phi\left(w_i^A, w_i^B, y_i\right) = \left(y_i^A, y_i^B\right)'$ *such that* $y_i^A + y_i^B = y_i$.

The sharing rule concept allows us to model household behaviour as a two-stage budgeting process. After dividing total nonlabour income in the first stage, each individual I $(I = A, B)$ faces a maximization problem that is formally similar to the unitary household decision problem, viz.:

$$\max_{c_i^I, l_i^I} v^I\left(c_i^I, l_i^I\right) \text{ subject to } c_i^I + w_i^I l_i^I \leq y_i^I + w_i^I T.$$

Chiappori (1992) demonstrated that the resulting household allocation is always Pareto efficient (and that each Pareto efficient allocation can be represented by such a two-stage budgeting process).

[2] The analysis is in fact also applicable to individual *caring preferences*, which can be represented by a utility function of the form $f^I\left(v^A\left(c^A, l^A\right), v^B\left(c^B, l^B\right)\right)$ $(I = A, B)$; see Chiappori (1992) for a detailed discussion.

[3] Evidently, collective and unitary models are the same for one-person households.

It turns out that this alternative interpretation of Pareto efficient household behaviour is particularly convenient within the nonparametric context, as it entails the same kind of GARP tests as for the unitary model. Indeed, if we knew private consumption for each observation (c_i^A and c_i^B), then we could immediately check consistency of the observed set S by using the standard GARP tests at the level of the *household members*. In practice, however, we do *not* observe the intrahousehold allocation of total consumption, and, hence, we obtain the following condition for the collective model (see also Chiappori, 1988):

Definition 6 *The observed set S is consistent with* collective rationalization with egoistic agents *if there exist n pairs of real numbers $\left(c_i^A, c_i^B\right)'$ such that for all $i = 1, ..., n$:*
$c_i^A + c_i^B = c_i$,
$c_i^A, c_i^B \geq 0$,
$c_i^A + c_i^B + w_i^A l_i^A + w_i^B l_i^B \leq y_i + w_i^A T + w_i^B T$
and
GARP is satisfied at the individual level ($I = A, B$)*:*
$\forall i, j \in \{1, ..., n\}$, *if* $\mathbf{l}_i^I R \mathbf{l}_j^I$ *then* $\mathbf{w}_j^{I\prime} \mathbf{l}_j^I \leq \mathbf{w}_j^{I\prime} \mathbf{l}_i^I$.

This condition constitutes the natural counterpart of the unitary GARP test. Indeed, given that the intrahousehold consumption allocation is not observed, we only need that there exists *at least one feasible* allocation entailing *individual* labour supply data $\{(c_i^I, l_i^I, w_i^I, y_i^I = c_i^I - w_i^I \ell_i^I), i = 1, ..., n\}$ that are consistent with GARP for *both* individuals.

The above exposition makes clear that we can directly apply the same GARP test as in the unitary model for each household member, conditional upon the sharing rule. Importantly, as shown by Chiappori (1988), the resulting collective test is not nested in the unitary test discussed above; consistency with the unitary GARP condition does not necessarily imply consistency with the collective GARP condition, and vice versa.[4]

Snyder (2000) introduced an ‘all-or-nothing’ nonparametric test for the collective model.[5] In that test, either data satisfy collective rationality or they do not. We follow a different approach, induced by our specific focus on the goodness-of-fit of the alternative behavioural models (see further; Section 2.3). Our starting point is that the collective rationalization test boils down to standard GARP tests conditional upon an intrahousehold consumption allocation (c_i^A and c_i^B).

[4] Note that Chiappori (1988) derived his results for a stronger consistency axiom than GARP, which he calls ‘Strong SARP’ (see also Chiappori and Rochet, 1987). Following Varian (1982), we start from the more general concept GARP, which suffices for most purposes of empirical nonparametric analysis. In addition, it is easy to show that our goodness-of-fit and power analysis in the next section remains unaffected when starting from Strong SARP instead of GARP.

[5] In her analysis, Snyder restricts attention to the case n=2, while we consider the more general case; e.g., in our application n=281 (see Section 3.1).

We propose a two-step procedure. In the first step, we test GARP for alternative 'uniform' allocation rules, i.e., rules that are simultaneously applied to all households in the sample. We then select the allocation rule with the highest number of individual (male and female) household members passing the associated GARP tests. In our application, we consider three such uniform rules: $\frac{c_i^A}{c_i} = \frac{2}{5}$, $\frac{c_i^A}{c_i} = \frac{1}{2}$ and $\frac{c_i^A}{c_i} = \frac{3}{5}$ for household member A; correspondingly, $\frac{c_i^B}{c_i} = 1 - \frac{c_i^A}{c_i}$ for household member B. The best rule in terms of the aforementioned selection criterion turns out to be $\frac{c_i^A}{c_i} = \frac{c_i^B}{c_i} = \frac{1}{2}$.

In the second step, we account for the possibility that different households may be characterized by other allocation rules, which (moderately) deviate from the selected uniform rule. Specifically, given our starting point ($\frac{c_i^A}{c_i} = \frac{c_i^B}{c_i} = \frac{1}{2}$), we randomly draw 2000 combinations of n consumption shares from a normal distribution with a cumulative probability of 95% for the values between 45% and 55%, i.e., $P\left(0.45 < \frac{c_i^A}{c_i} < 0.55\right) = 0.95$.[6] Like in the first step, we retain the combination of shares that is associated with the highest number of individual household members passing GARP; this combination is used for comparing the empirical performance of the collective model with that of the unitary model.

As a final note, we point out that this approach does not guarantee the generally most favourable treatment of the collective model: to ensure computational tractability, our procedure restricts attention to a limited number of possible combinations of intrahousehold allocations; there may well exist other, non-investigated combinations that are associated with an even higher number of individuals consistent with GARP.[7] We can therefore conclude that our empirical (goodness-of-fit and power) analysis implicitly gives the 'benefit of the doubt' to the standard unitary model and, in that sense, we can call it 'conservative'.

2.3 Empirical performance: goodness-of-fit

The consistency tests reviewed above are 'sharp' tests; they only tell us whether observations are *exactly* optimizing in terms of the behavioural model that is under evaluation. However, as argued by Varian (1990), *exact* optimization is not a very interesting hypothesis. Rather, we want to know whether the behavioural model under study provides a *reasonable* way to describe observed behaviour; for most purposes, 'nearly optimizing behaviour' is just as good as 'optimizing' behaviour. Varian's argument is all the more valid in the context of comparing theoretical behavioural models: we are primarily interested in the

[6] We thus assume individuals' consumption shares to be distributed according to $N\left(\mu, \sigma^2\right)$, with $\mu = 0.5$ and $\sigma = 0.0255102$. We have also experimented with larger σ values, but the corresponding results suggest a worse fit of the collective model.

[7] From that point of view, a 'better' procedure may consist of a grid search which explores all combinations of intrahousehold allocations between, e.g., $\frac{c_i^A}{c_i} = 0.1$ and $\frac{c_i^A}{c_i} = 0.9$ with interval 0.01. However, this alternative is computationally extremely cumbersome and therefore not easy to implement in practice; e.g., the example implies checking 81^n different combinations.

extent to which one model 'fits' the observed data better than the other model. Therefore, our following assessment will be based on measures of *goodness-of-fit* rather than on the mere consistency tests as such.

Our goodness-of-fit measure is the 'improved violation index' (or 'efficiency index') proposed by Varian (1993; see also Cox, 1997), which indicates the *degree* to which the data are 'optimizing' (or 'efficient') in the sense of the evaluated behavioural model.[8] More specifically, this index gives for each observation the minimal perturbation in the associated budget set that will satisfy the optimization (or 'efficiency') condition; i.e., the perturbation that guarantees consistency of the observed set S with GARP. (Recall from our previous discussion that the unitary model and the collective model entail formally the same GARP tests, so that the following discussion directly applies to both approaches.) The focus on budget shifts links up with Definition 4, which states that consistency of the set S with GARP requires all observations to be expenditure minimizing over their revealed preferred set.

We refer to Varian (1993) and Cox (1997) for in-depth formal discussions of Varian's improved violation index, and restrict to a graphical illustration in the current study. To keep the exposition simple, we only illustrate the individuals' case. Figure 1 contains 2 leisure(l)-consumption(c) observations, which we refer to as 1 and 2. The relative prices for the consumption bundles 1 and 2 correspond to the slopes of the respective budget hyperplanes C1 and C2. Obviously, both observations imply a violation of GARP: for observation 1, the ratio between minimal expenditure (over the revealed preferred set, which includes both 1 and 2) and actual expenditure equals 01'/01, i.e. the (relative) radial distance between the budget hyperplanes C1' and C1; similarly, observation 2 violates GARP by the fraction 02'/02, i.e. the radial distance between C2' and C2.

The basic concept in Varian's procedure is the 'violation index'. For each observation, this index captures the degree to which actual expenditure exceeds minimal expenditure, as defined over the revealed preferred set. It varies between 0 and 1 by construction; a value of 1 suggests behaviour that is consistent with the optimization hypothesis under study. In our example, the proportions 01'/01 and 02'/02 are the violation index values associated with respectively observation 1 and observation 2.

In the Varian (1993) terminology, both observations in Figure 1 are involved in a 'revealed preference cycle'. Varian then proposes a procedure that identifies the minimal expenditure perturbations needed to 'break' this cycle. The central idea behind the procedure is that a cycle can often be eliminated by perturbing just one of the budget hyperplanes involved in the cycle; it is not necessary to shift the budget hyperplanes of all consumption bundles. Specifically, Varian's procedure starts from the basic violation index to construct an *improved* violation index for each observation. This improved index captures the *minimal* budget hyperplane perturbations associated with the respective consumption

[8] Cox (1997) shows that Varian's improved violation index can also be interpreted as correcting the sharp test procedure for over- or underreporting in the expenditure data.

bundles to obtain consistency with GARP.

In our graphical example, it suffices to shift the budget hyperplane C1 by a (positive) factor that is strictly below the associated violation index (e.g., e*(01'/01)<(01'/01), with e close to unity). A test for optimizing behaviour that is weaker than the orginal 'sharp' test then multiplies the original expenditure level of observation 1 by that factor, while leaving the expenditure level of observation 2 unaltered. It turns out that we cannot reject GARP for these newly constructed expenditure values: observation 1 is no longer revealed preferred to observation 2 (and, hence, observation 2 is consistent with GARP by construction), and is itself expenditure minimizing over its revealed preferred set (which includes both 1 and 2). Notice that 01'/01 is closer to unity than 02'/02, so that shifting C1 (by e*(01'/01)) is less 'drastic' than shifting C2 (by, e.g., e*(02'/02)). Hence, Varian's procedure selects e*(01'/01) and 1 (and not 1 and e*(02'/02)) as the *improved* violation index values corresponding to observations 1 and 2, respectively.

Of course, in the general case with multiple observations, more than two consumption bundles are often involved in a revealed preference cycle. For this case, Varian proposes an iterative algorithm for computing improved violation index values. We employ this algorithm in our empirical application.[9]

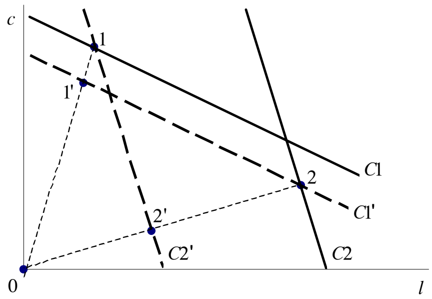

Figure 1: The improved violation index for individuals

2.4 Empirical performance: power

A full empirical assessment should contain more than only a goodness-of-fit analysis of the alternative models; we believe it important to additionally account for the *power* of each model. Indeed, favourable goodness-of-fit results, indicating few violations of the behavioural restrictions, have little meaning if the behavioural implications are hardly restrictive; i.e., optimizing behaviour can hardly be rejected.

The construction of our nonparametric power measure follows Bronars (1987; see also Cox, 1997). That is, we use a randomization procedure to construct the power measure. This procedure is based on Becker's (1962) notion of *irrational* behaviour, which states that a consumer chooses consumption bundles

[9] In this application, we set the factor e (see our above example) equal to 0.9999999999.

randomly from his budget set. More specifically, Beckerian irrational behaviour means that the consumer chooses consumption bundles from a uniform distribution across all bundles in the budget hyperplane. Bronars' power measure then captures the probability of rejecting the null hypothesis of optimizing (or 'rational') behaviour in case of such irrational (or 'random') behaviour.[10]

We quantify power of the GARP test associated with each model as follows. Firstly, we simulate irrational/random behaviour for each observation in the set S. That is, for the different (leisure and consumption) commodities we draw random budget shares (of at least 1%, summing to 100%) from a continuous uniform distribution. The generated budget shares are then multiplied by observed total expenditure and divided by the actual price of each commodity, to obtain the random consumption of each commodity. Subsequently, we check consistency with the GARP condition for each observation, based on the simulated ('irrational') quantity bundles and actual prices. In our empirical application, we repeat this procedure 1000 times. For each observation, the proportion of rejections of GARP (over these 1000 replications) gives the probability of detecting irrational behaviour of that observation, given random behaviour of the other observations.

Hence, for each model that we evaluate we measure power in each element of the observed set S. This practice contrasts with Bronars (1987) and Cox (1997), who provide overall power measures that are based on the entire sample. Their measures reveal the probability that random behaviour of at least one observation in the sample is detected.

In our opinion, evaluating power at the level of individual observations is more informative. For example, it provides a much more detailed insight into the extent to which the different observations *can* cause rejection of the model under study; we believe that there is a stronger case for a model that has high power in many observations than for a model with high power in only a few observations. Also, an observation-specific power measure naturally links up with our observation-specific goodness-of-fit measure; persistently high goodness-of-fit values for a given sample of observations are all the more convincing evidence in favour of a particular behavioural model if they are complemented with generally high power values for the same sample.

3 Application

3.1 Data and methodological issues

Our data are drawn from the 1992 and 1997 waves of the Socio-Economic Panel (SEP) of the Center for Social Policy (University of Antwerp). Specifically, we focus on three subsamples: female singles, male singles and couples. The first two subsamples consist of female and male singles that meet the following cri-

[10] As discussed by Bronars, that probability directly depends on the number of budget set intersections associated with the different consumption bundles under study; if there are no intersections, then irrational behaviour cannot be detected.

teria: no children, aged between 25 and 55 and employed. The third subsample consists of (de-facto) couples, where the household members meet the same criteria as the selected singles. To minimize the impact of measurement error, we have trimmed out from each subsample those households that include a (female/male) member with a wage that lies above the 97.5 percentile or below the 2.5 percentile of the empirical (female/male) wage distribution. This yields samples of 123 single females, 173 single males and 281 couples.

Cox (1997) and Snyder (2000) also conduct nonparametric tests of labour supply behaviour on micro-data (individuals and households). They test consistency with GARP of time-series data and, hence, they exclude preference variation over time. Our analysis deviates slightly in that we assume constant preferences in each cross-section subsample (female singles, male singles and couples); in each subsample, all observations correspond to the same preferences but to different price regimes.

Our motivation for this particular preference homogeneity assumption is threefold. Firstly, the SEP was subject to substantial attrition between 1992 and 1997: because many new households entered the data set in 1997, only a small number of households were observed in both waves of the SEP; there are too few households with two consecutive observations for robust nonparametric testing based on time-series data. Secondly, our selection criteria ensure relatively homogeneous subsamples, which makes that our equal preference assumption does not seem overly strong.[11] Finally, and importantly, recall that we focus on goodness-of-fit measures in our following analysis. Obviously, this practice anticipates (slight) preference variation over households. As a matter of fact, we believe that our preference homogeneity assumption, which is indispensable for a meaningful application, directly calls for a nonparametric goodness-of-fit analysis rather than a mere examination of the results obtained from the 'sharp' consistency tests discussed in Sections 2.1 and 2.2.[12]

3.2 Singles versus couples

Figure 2 presents the cumulative distribution functions (c.d.f.'s) of the goodness-of-fit measures (i.e. the improved violation or efficiency indexes) associated with the unitary model for female singles, male singles and couples.[13] When restricting to the 'sharp' GARP condition, we would conclude rejection for all three subsamples; relatively few observations have an index value that equals 100%. Recalling our earlier discussion, however, this result should not be very surprising given our assumption of equal preferences over all observations in

[11] Compare, e.g., with Famulari (1995). She analyses consistency of observed behaviour with GARP (in a unitary framework) for homogenous subgroups of households that are identified on the basis of similar selection criteria.

[12] In this sense, our goodness-of-fit measures have an interpretation that is comparable to that of the unobserved error terms in parametric regressions, where similar households are assumed to have 'more or less' the same preferences.

[13] For expositional convenience, the c.d.f.'s have been cut off at the 91% efficiency level since no observation has a violation index below that figure. We also explicitly distinguish between indexes that are equal to 1 and those that are less than 1.

each subsample. It seems more meaningful to look at the *entire* distribution of the goodness-of-fit measure.

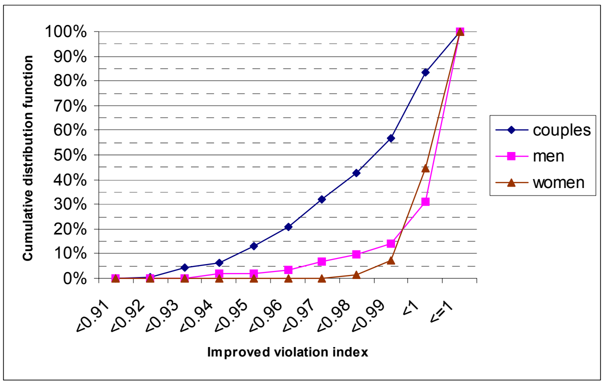

Figure 2: Unitary model singles and couples: cumulative distribution function of improved violation index

When considering the c.d.f.'s more closely, we observe important differences between couples and singles. Firstly, we find that 55% of the female singles and 69% of the male singles are fully efficient in terms of the improved violation index, as opposed to only 17% of the couples. Secondly, and more importantly, the index values of couples are generally below those of singles; the couples' distribution is stochastically dominated by the two (female and male) singles' distributions. One-tailed Kolmogorov-Smirnov tests confirm this overall picture: the null hypothesis of equal distributions of couples on the one hand and male and female singles on the other hand is rejected (at any conventional significance level) in favour of the alternative hypothesis that the couples' index systematically lies below the respective singles' indexes; see Table 1.

A final comparison of these goodness-of-fit results is based on Varian (1990). When introducing the nonparametric goodness-of-fit idea, Varian (p.129) suggests the 'magic number' of significance tests, 5%, as 'probably a reasonable choice' for evaluating consistency of observed behaviour with the model under study. This suggests a nonparametric '95% (= 100%-5%) test' for consistency with the optimization hypothesis. Hence, it seems worthwhile to compare the number of observations of each subsample that pass this test. Once more, such comparison reveals that the unitary model fits the singles' data much better than the couples' data: all female singles and 98% of the male singles are at least 95% efficient, in contrast to only 87% of the couples; see Figure 2.

As discussed before, a comparison that is solely based on goodness-of-fit can be misleading: goodness-of-fit differences may be caused by power differences. For the sake of completeness (and fairness), we therefore complement our goodness-of-fit analysis with a power analysis.

The c.d.f.'s of the individually calculated power indexes for single females, single males and couples are shown in Figure 3.[14] This figure reveals high power

[14] In contrast to Figure 2, Figure 3 presents the whole c.d.f. The reason is that a few

for most observations: 96% of the couples, 92% of the male singles and 89% of the female singles have a power index value that exceeds 95%; for these observations, irrational/random behaviour will be detected with a probability of at least 95%. More generally, while the overall power for couples appears to be slightly higher than for female and -to a somewhat lesser extent- male singles, Figure 3 suggests that the differences remain marginal. This impression is confirmed by one-tailed Kolmogorov-Smirnov tests: we cannot reject (at the 5% significance level) equality of the c.d.f.'s in favour of the alternative hypothesis that the power index values for (female and male) singles are lower than those for couples; see Table 1.

We conclude that the relatively poor performance of the unitary model for describing observed couples' behaviour (when compared to singles' behaviour) can hardly be attributed solely to higher power of the model for the associated couples' consistency tests. In our opinion, these findings strongly question the harmless nature of the aggregation assumptions in the unitary approach to modelling couples' behaviour.

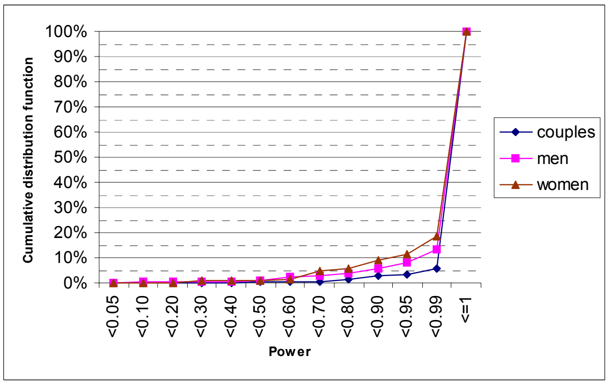

Figure 3: Unitary model singles and couples: cumulative distribution function of power

Table 1: Differences between singles and couples

	Imp. viol. index	**Power index**
Single women vs. couples	0.000	0.055
Single men vs. couples	0.000	0.290

Entries show the probability that the null hypothesis of equal distribution is true, as computed on the basis of a one-tailed Kolmogorov-Smirnov test; we compare the distributions of the improved violation index and the power index for couples with the respective distributions for single women and single men.

observations have very low power indexes.

3.3 Unitary versus collective model

Our previous findings cast doubts on the usefulness of the unitary model for analyzing couples' behaviour. As a natural next step, we now investigate whether the collective approach provides a better alternative for modelling couples' behaviour, by comparing its empirical performance with that of the unitary model. Like before, our unitary results refer to GARP tests at the *aggregate household* level. By contrast, our collective results are obtained from applying GARP tests to the *individual members* of each couple, hereby using the intrahousehold allocations obtained by the procedure described in Section 2.

Figure 4 presents the c.d.f.'s of the goodness-of-fit measure for couples (in the unitary model) and female and male household members (in the collective model). In line with our earlier results, substantially more individuals than (aggregate) households behave consistently with the utility maximization hypothesis: 38% of the men and 35% of the women are 100% efficient, while only 17% of the couples attain an improved violation index of 100%. In fact, Figure 4 reveals a picture that is roughly similar to that in Figure 2: the (unitary) couples' distribution is stochastically dominated by the (collective) distributions of the male and female household members. The Kolmogorov-Smirnov test results in Table 2 provide further evidence in support of the collective model: the null hypothesis of equal c.d.f.'s is strongly rejected in favour of the alternative hypothesis that the couples' improved violation index systematically lies below that for women and men in the collective model. Finally, also the 95% consistency tests favour the collective model: from Figure 4 we observe that in the collective model all women and 99% of the men are at least 95% efficient, while this applies to only 87% of the couples in the unitary model.

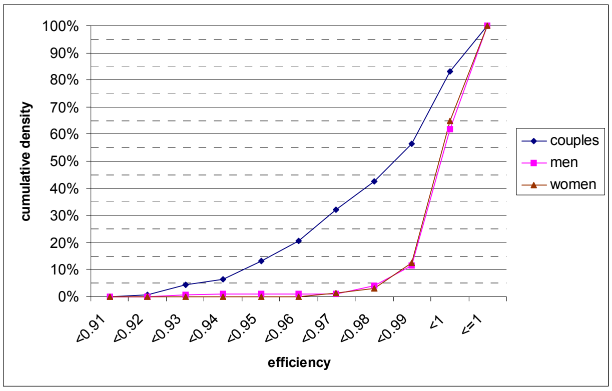

Figure 4: Unitary versus collective model couples: cumulative distribution function of improved violation index

Again, we complement this goodness-of-fit analysis with a power analysis. Our power results persistently indicate that the better fit of the collective model is not due to lower power; the hypothesis that there are no power differences is even better supported than in the previous section, where we observed slightly (although not significantly) higher power of the unitary model in the couples' case than in the singles' case. For example, Figure 5 clearly shows that the distribution of the power indexes is practically the same for couples (in the unitary model) and individuals (in the collective model). This observation is formalized in Table 2: one-tailed Kolmogorov-Smirnov tests reveal that equality of the c.d.f.'s of the power indexes cannot be rejected at any conventional significance level. Moreover, the power indexes are generally high: 96% of the couples (in the unitary model), 96% of the females and 95% of the males (in the collective model) have a power index that amounts to at least 95%; see Figure 5.

In our opinion, these results provide strong enough evidence to argue that the collective approach performs significantly better than the unitary approach in the modelling of couples' labour supply behaviour. In fact, this argument becomes all the more convincing when taking into account our rather rudimentary procedure to model the intrahousehold allocation; more refined allocation identification procedures (e.g., based on a grid search; cf. supra) can only benefit the relative performance of the collective model.

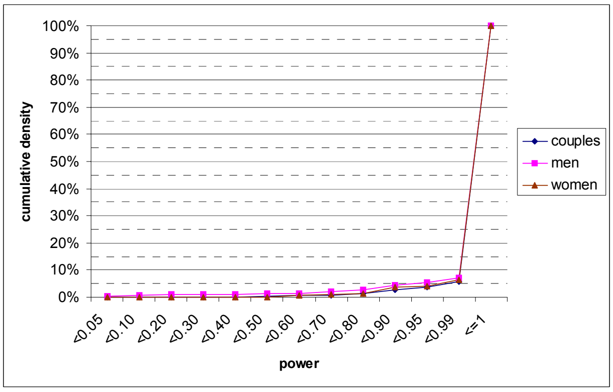

Figure 5: Unitary versus collective model couples: cumulative distribution function of power

Table 2: Differences between the unitary model and the collective model

	Imp. viol. index	**Power index**
Single women vs. couples	0.000	0.934
Single men vs. couples	0.000	0.991

Entries show the probability that the null hypothesis of equal distribution is true, as computed on the basis of a one-tailed Kolmogorov-Smirnov test; we compare the distributions of the improved violation index and the power index for couples in the unitary model with the respective distributions for women and men in the collective model.

4 Conclusion

We have compared the empirical performance of the standard unitary approach to modelling household labour supply behaviour with that of the more recently developed collective approach. Specifically, we quantified empirical performance in terms of well-defined goodness-of-fit measures, which tell us to what extent the behavioural model explains observed behaviour. Our analysis deviates from the mainstream literature in that we employ nonparametric tools to quantify the performance of the different models; nonparametric analysis avoids the possibly distortive effects of an ill-specified functional form for the preferences and/or the intrahousehold bargaining process. We have complemented our goodness-of-fit analysis with an in-depth power analysis; power measures allow for evaluating the extent to which favourable goodness-of-fit results may be due to low discriminatory power of the associated consistency tests.

Our findings strongly suggest using the collective model for analyzing the behaviour of households consisting of multiple individuals. For example, we found that the unitary model performs significantly worse when applied to couples than when applied to singles. As these results could not be explained by significant power differences, we conclude that they reflect violations of the preference aggregation assumptions that underlie the unitary approach, i.e., that a household behaves as a single decision maker. Indirectly, this provides an argument in favour of the collective model, as the latter explicitly recognizes that multi-person households consist of several individuals with own (possibly diverging) preferences.

Direct comparison of the collective model with the unitary model provided additional evidence to support the use of the collective model: the collective model fits observed couples' behaviour much better than the unitary model. Again, this significant difference cannot be attributed to power differences. Hence, our findings do not only indicate that the unitary approach is too restrictive for modelling couples' behaviour, but also that the collective model constitutes a more promising alternative.

In a way, our nonparametric analysis reproduces the parametric results of Browning and Chiappori (1998). By adopting an explicit nonparametric orientation, we counter the possible criticism on those latter results that they primarily reflect the impact of the (non-verifiable) functional form that is employed. In our opinion, the main strength of our analysis is precisely that it tests the validity of the different behavioural models directly on the observed data. More generally, we believe that our study illustrates that nonparametric analysis is particularly useful for evaluating alternative, 'competing' behavioural models, especially since they avoid the 'black box' of the functional form.

We see at least two avenues for further research within this nonparametric orientation. Firstly, a more advanced modelling of the intrahousehold consumption allocation can be pursued. In our opinion, the two-stage procedure used in the current paper can provide a useful starting point towards a (computationally tractable) procedure based on goodness-of-fit criteria. A second, closely related research topic concerns the issue of recovering the sharing rule and the individual preferences within a collective model; compare with Varian (1982), who addresses similar questions within the unitary framework. Indeed, as emphasized at various occasions in the existing literature (e.g., Chiappori, 1992), detailed knowledge of these concepts can be very informative for welfare comparisons.

References

[1] Afriat, S. (1967), “The construction of utility functions from expenditure data”, *International Economic Review*, 8, 67-77.

[2] Afriat, S. (1972), “Efficiency estimation of production functions”, *International Economic Review*, 13, 568-598.

[3] Becker, G. (1962), “Irrational behaviour and economic theory”, *Journal of Political Economy*, 70, 1-13.

[4] Blundell, R. (1988), “Consumer behaviour: theory and empirical evidence - A survey”, *Economic Journal*, 98, 16-65.

[5] Blundell, R., M. Browning and I. Crawford (2001), “Nonparametric Engel curves and revealed preferences”, forthcoming in *Econometrica.*

[6] Bronars, S. (1987), “The power of nonparametric tests of preference maximization”, *Econometrica*, 55, 693-698.

[7] Browning, M., F. Bourguignon, P.-A. Chiappori and V. Lechene (1994), “Income and outcomes: a structural model of intrahousehold allocation”, *Journal of Political Economy*, 102, 1067-1096.

[8] Browning, M. and P.-A. Chiappori (1998), “Efficient intra-household allocations: a general characterization and empirical tests”, *Econometrica*, 66, 1241-1278.

[9] Chiappori, P.-A. (1988), “Rational household labor supply”, *Econometrica*, 56, 63-89.

[10] Chiappori, P.-A. (1992), “Collective labor supply and welfare”, *Journal of Political Economy*, 100, 437-467.

[11] Chiappori, P.-A., B. Fortin and G. Lacroix (2002), “Marriage market, divorce legislation and household labor supply”, *Journal of Political Economy*, 110, 37-72.

[12] Chiappori, P.-A. and J.-C. Rochet (1987), “Revealed preferences and differentiable demand”, *Econometrica*, 55, 687-691.

[13] Cox, J. (1997), “On testing the utility hypothesis”, *Economic Journal*, 107, 1054-1078.

[14] Famulari, M. (1995), “A household-based, nonparametric test of demand theory”, *Review of Economics and Statistics*, 77, 372-382.

[15] Fortin, B. and G. Lacroix (1997), “A test of the unitary and collective models of household labour supply”, *Economic Journal*, 107, 933-955.

[16] Snyder, S. (2000), “Nonparametric testable restrictions of household behaviour”, *Southern Economic Journal*, 67, 171-185.

[17] Varian, H. (1982), “The nonparametric approach to demand analysis”, *Econometrica*, 50, 945-973.

[18] Varian, H. (1990), “Goodness-of-fit in optimizing models”, *Journal of Econometrics*, 46, 125-140.

[19] Varian, H. (1993), “Goodness-of-fit for revealed preference tests”, Mimeo, Ann Arbor, University of Michigan.