Nonparametric analysis of household labour supply:

Goodness-of-fit and power of the unitary and the collective model*

Laurens Cherchye† and Frederic Vermeulen‡

August 2004

Abstract

We compare the empirical performance of the unitary and the collective approach to modelling observed labour supply behaviour. A nonparametric analysis is conducted, which avoids the distortive impact of an erroneously specified functional form for the preferences and/or the intrahousehold bargaining process. Our analysis 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.

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

JEL-classification: C14, D12, J22.

*We thank two anonymous referees for helpful comments and suggestions. We are also grateful to Jan Annaert, Denis Beninger, Martin Browning, Pierre-André Chiappori, André Decoster, David Margolis, Erik Schokkaert and Dirk Van de gaer, as well as seminar participants in Copenhagen, Ghent, Leuven, Tilburg, ESEM 2003 Stockholm and JMA 2004 Lille for useful remarks.

†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, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: frederic.vermeulen@uvt.nl. Frederic Vermeulen acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00235, (AGE).

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 *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., Fortin and Lacroix, 1997, and Browning and Chiappori, 1998).

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 unitary restrictions; e.g., Browning *et al.* (1994), Fortin and Lacroix (1997), Browning and Chiappori (1998), Chiappori *et al.* (2002) and Vermeulen (2004).

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 solely based on revealed preference theory, which makes them particularly attractive for testing consistency of the data with theoretical behavioural models.

This directly suggests using nonparametric testing tools for comparing the empirical performance of the unitary and collective models. However, to the best of our knowledge, an in-depth nonparametric comparison has not yet been carried out. This paper wants to fill that gap, 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 (nonparametric) empirical performance measures: *goodness-of-fit* measures and *power* measures. We indeed believe that a fair comparison of the two behavioural models under study should complement a goodness-of-fit analysis with a power analysis: favourable goodness-of-fit results, indicating few violations of the behavioural restrictions, have little meaning if the behavioural implications have low power, i.e., optimizing behaviour can hardly be rejected.

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. 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 a better insight into which of the models does the better job in describing multi-person household consumption behaviour.

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.

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 (M and F). 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 ($I = M, F$) 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 nonlabour income and consumption. Finally, we represent the set of all observations by $S = \left\{\left(c_i, l_i^M, l_i^F, w_i^M, w_i^F, y_i\right), i = 1, ..., n\right\}$.

Within the unitary model, the decision problem of each household i boils down to maximizing a nonsatiated utility function $v\left(c_i, l_i^M, l_i^F\right)$ subject to the budget constraint $c_i + w_i^M l_i^M + w_i^F l_i^F \leq y_i + w_i^M T + w_i^F T$; 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 function v that *rationalizes* the household data, i.e., for all $i \in \{1, ..., n\}$ the value $v\left(c_i, l_i^M, l_i^F\right)$ equals:

$$\max_{j \in \{1,...,n\}} v\left(c_j, l_j^M, l_j^F\right) \text{ s.t. } c_j + w_j^M l_j^M + w_j^F l_j^F \leq y_i + w_i^M T + w_i^F T. \quad (1)$$

Varian (1982) has demonstrated that such a data rationalizing utility function exists 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 $\left(1, w^M, w^F\right)' = \mathbf{w}$ and $\left(c, l^M, l^F\right)' = \mathbf{l}$):

Definition 1 *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 2 *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 definition expresses the idea that observation j 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 is easily tested: we first identify the set RP_j and consequently check the expenditure minimization condition. See Varian (1982; p. 949) for an efficient algorithm.

2.2 Testing the collective model

We focus on the collective consumption model with *egoistic* preferences; preferences only depend on own consumption and leisure (Chiappori, 1988). Moreover, we assume that there is no public consumption in the household.[1] 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.

In the following, 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 $\mathbf{l}_i^I$ $(I = M, F)$. Using this, we consider the case where each couple i is characterized by a pair of (nonsatiated) utility functions, $v^M\left(c_i^M, l_i^M\right)$ and $v^F\left(c_i^F, l_i^F\right)$, *and* a sharing rule $\phi\left(w_i^M, w_i^F, y_i\right)$ that determines the distribution of the household's nonlabour income y_i over the household members (see Chiappori, 1988):

Definition 3 *A* sharing rule ϕ *is a function which maps the vector* $\left(w_i^M, w_i^F, y_i\right)'$ *to* $\phi\left(w_i^M, w_i^F, y_i\right) = \left(y_i^M, y_i^F\right)'$ *such that* $y_i^M + y_i^F = y_i$.

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 $(I = M, F)$ of the couple i faces the maximization problem:

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

which is formally similar to the unitary household decision problem; see (1). Chiappori (1992) has demonstrated that the resulting household allocation is always Pareto efficient.

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: if we knew private consumption for each observation (c_i^M and c_i^F), then we could immediately check consistency

[1] The analysis is in fact also applicable to individual *caring preferences*, which can be represented by a utility function of the form $f^I\left(v^M\left(c^M, l^F\right), v^F\left(c^F, l^F\right)\right)$ $(I = M, F)$; see Chiappori (1992) for a detailed discussion. Cherchye, De Rock and Vermeulen (2004) provide a nonparametric characterization of a general collective model with public goods and externalities. Given the current paper's objective, we focus on a rather simple collective model, which can be considered as a direct generalization of the unitary model.

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. This entails the following empirical condition for the collective model (see also Chiappori, 1988):

Definition 4 *The observed set S is consistent with* collective rationalization with egoistic agents *if there exist n pairs of real numbers $\left(c_i^M, c_i^F\right)'$ such that for all $i = 1, ..., n$:*
$c_i^M + c_i^F = c_i,$
$c_i^M, c_i^F \geq 0,$
$c_i^M + c_i^F + w_i^M l_i^M + w_i^F l_i^F \leq y_i + w_i^M T + w_i^F T$
and
GARP is satisfied at the individual level ($I = M, F$)*:*
$\forall i, j \in \{1, ..., n\}$, *if* $\mathbf{l}_i^I R \mathbf{l}_j^I$ *then* $\mathbf{w}_j^{I\prime} \mathbf{l}_j^I \leq \mathbf{w}_j^{I\prime} \mathbf{l}_i^I$.

Thus, 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 $\{(c_i^I, l_i^I, w_i^I, y_i^I = c_i^I - w_i^I \ell_i^I); i = 1, ..., n; I = M, F\}$ that are consistent with GARP for *both* individuals.

Snyder (2000) introduced an 'all-or-nothing' nonparametric test for the collective model.[2] 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. Our starting point is that the collective rationalization test boils down to standard GARP tests conditional upon an intrahousehold consumption allocation (c_i^M and c_i^F). Specifically, we impute (unobserved) member-specific private consumption amounts by exploiting a systematic finding in parametric studies of collective labour supply, namely the apparently positive correlation between the male/female member's share of total nonlabour income and the corresponding individual wage (e.g., Chiappori *et al.*, 2002, and Vermeulen, 2004).

Using this, our nonparametric testing exercise considers the following pair of distributions for the female consumption share s_i^F ($= c_i^F / c_i$; the corresponding male share equals $1 - s_i^F$): the first distribution has mean 0.45 and a cumulative probability of 95% for the values between 0.40 and 0.50; the second distribution has mean 0.55 and a cumulative probability of 95% for the values between 0.50 and 0.60. From these distributions, we draw 1000 combinations of s_i^F values ($i = 1, ..., n$): if $w_i^M \geq w_i^F$ ($w_i^F > w_i^M$) then s_i^F is drawn from the first (second) distribution. We subsequently select the combination of shares with the highest number of individual (male and female) household members passing the associated GARP tests. This combination is used for comparing the empirical performance of the collective model with that of the unitary model.[3]

As a final note, we emphasize that this approach does not guarantee the 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

[2] 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).

[3] We have also experimented with alternative means for the above normal distributions (including a rule where both distributions have mean 0.50). But this did not yield a higher number of (male and female) household members passing the GARP tests.

individuals consistent with GARP. We can therefore state that our empirical analysis implicitly gives the 'benefit of the doubt' to the unitary model.

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 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*.

Our goodness-of-fit measure is the 'improved violation index' (or 'efficiency index') proposed by Varian (1993; based on Afriat, 1973; see also Cox, 1997), which indicates the *degree* to which the data are 'optimizing' (or 'efficient') in the sense of the evaluated behavioural model. More specifically, this index gives for each observation the minimal perturbation of the expenditure level that guarantees consistency of the observed set S with GARP. See Varian (1993) and Cox (1997) for in-depth formal discussions of this goodness-of-fit measure.

2.4 Empirical performance: power

We compute four different power measures. A first distinction relates to the consumption data that is used. The first dataset (used for the measures *Power1a* and *Power1b*) consists of the original consumption and leisure data for the unitary model and the combination of observed and partly imputed data for the collective model. The second dataset (used for the measures *Power2a* and *Power2b*) multiplies the original expenditure level for each observation with the corresponding improved violation index value (at the household level for the unitary model and at the individual level for the collective model); this anticipates the question to what extent the necessary data perturbation for obtaining GARP consistency (captured by the improved violation index) would effectively impact on the power estimates.

For each dataset we compute two types of power measures proposed by Bronars (1987). The distinction between the two types essentially pertains to the 'mimicking' of irrational behaviour *à la* Becker (1962), by means of a specific randomization procedure for constructing irrational consumption bundles; each power measure then captures the probability of detecting that irrational behaviour. The first type of measures, labelled *Power1a* (*Power2a*) for the first (second) dataset, uses Bronars' first algorithm; the second type of measures, labelled *Power1b* (*Power2b*) for the first (second) dataset, uses Bronars' second algorithm. We refer to Bronars for formal definitions, and restrict here to indicating that the first algorithm generally implies a greater probability than the second one of generating (irrational) consumption bundles near the 'corners' of the budget line. Or conversely, the second algorithm makes 'extreme' (irrational) behaviour less likely than the first one.

For a given data set and randomization procedure, the specific construction of the power measures first simulates irrational/random behaviour for each observation, and subsequently checks consistency with the GARP condition for each observation. In our empirical application, we repeat this procedure 200 times. The proportion of rejections of GARP (over these 200 replications) then gives the probability of detecting irrational behaviour of each observation, given random behaviour of the other observations.

Hence, for each behavioural model that we evaluate we measure power in *each* element of the observed set S. This practice contrasts with e.g. 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 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 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). 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 criteria: no children, aged between 25 and 55 and employed. The third subsample consists of (defacto) 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. The Appendix reports summary statistics for each subsample.

Cox (1997) and Snyder (2000) also conduct nonparametric tests of labour supply behaviour on micro-data. They test consistency with GARP on time-series data and, hence, they exclude preference variation over time. Our analysis deviates 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.[4] Finally, and importantly, recall that we focus on goodness-of-fit measures in our following analysis. Obviously, this practice anticipates some preference variation over households.

3.2 Singles versus couples

Figure 1 presents the cumulative distribution functions (c.d.f.'s) of the goodness-of-fit measures (i.e., the improved violation indexes, in ascending order) associated with the unitary model for female singles, male singles and couples.[5] 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%. We note that this result should not be very surprising in view of our preference homogeneity assumption. It seems more meaningful to look at the *entire* distribution of the goodness-of-fit measure.

[Figure 1 about here]

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, 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 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 in favour of the alternative hypothesis that the couples' index systematically lies below the respective singles' indexes; see Table 1.

As discussed before, it is recommendable to complement this goodness-of-fit analysis with a power analysis. Figure 2 presents the c.d.f.'s of the individually calculated *Power1a* indexes for single females, single males and couples.[6] This figure reveals high power 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 male singles, Figure 2 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.

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

[5] 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.

[6] In contrast to Figure 1, Figure 2 presents the whole c.d.f. The reason is that a few observations have very low power indexes.

We obtain similar results for the other three power measures.[7] First, the measure *Power2a* entails exactly the same qualitative conclusions as *Power1a*: power index values are generally very high, while equality of the c.d.f.'s for the three subsamples cannot be rejected. Finally, for the measures *Power1b* and *Power2b* we reject equality of the c.d.f.'s for singles and couples in favour of the alternative hypothesis that couples are associated with higher power index values than male and female singles. But also in this case the power index values remain generally very high for the different subsamples.

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 2 about here]

[Table 1 about here]

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 2.2.

Figure 3 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: 53% of the men and 58% of the women are 100% efficient, while only 17% of the couples attain an improved violation index value of 100%. In fact, Figure 3 reveals a picture that is roughly similar to that in Figure 1: 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.

[Figure 3 about here]

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 Section 3.2, where we observed slightly (although

[7] For the sake of brevity, we do not report the results here. But, of course, they are available from the authors upon request. This qualification equally applies for our discussion of the same power measures in Section 3.3.

not always significantly) higher power of the unitary model in the couples' case than in the singles' case. For example, Figure 4 clearly shows that the distribution of the *Power1a* values 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. 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%. Just like in Section 3.2, we have checked the sensitivity of these power results. Interestingly, the measures *Power2a*, *Power1b* and *Power2b* entail exactly the same qualitative conclusions as *Power1a*.

In our opinion, these results provide strong enough evidence to argue that the collective approach performs significantly better than the unitary approach for modelling couples' labour supply behaviour. In fact, this argument becomes all the more convincing when taking into account our rather rudimentary procedure to model the distribution of household consumption over the different household members; more refined allocation rules can only benefit the relative performance of the collective model.

[Figure 4 about here]

[Table 2 about here]

4 Conclusion

We have compared the empirical performance of the unitary model to describe household labour supply behaviour with that of the more recently developed collective model. Our findings strongly suggest using the collective model for analyzing the behaviour of households consisting of multiple individuals:
- First, we found that the unitary model performs significantly worse when applied to couples than when applied to singles. As these results can hardly be attributed to power differences, we conclude that they signal violations of the preference aggregation assumptions that underlie the unitary approach, i.e., that multi-person households behave as single decision makers.
- Second, and probably more importantly, a direct comparison of the collective model with the unitary model provided additional evidence to support the use of the collective model: it fits observed couples' behaviour much better than the unitary model. Again, this significant difference cannot be explained by power differences. Hence, our findings do not only indicate that the unitary approach is too restrictive for modelling the behaviour of multi-person households, but also that the collective model constitutes a more promising alternative.

References

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

[2] Afriat, S. (1973), "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method", *International Economic Review* 14, 460-472..

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

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

[5] 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.

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

[7] Cherchye, L., B. De Rock and F. Vermeulen (2004), "The collective model of household consumption: a nonparametric characterization", Mimeo, CentER, Tilburg.

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

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

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

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

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

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

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

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

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

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

[18] Vermeulen, F. (2004), "And the winner is... An empirical evaluation of unitary and collective labour supply models", forthcoming in *Empirical Economics*.

Appendix: Data

[Table 3 about here]

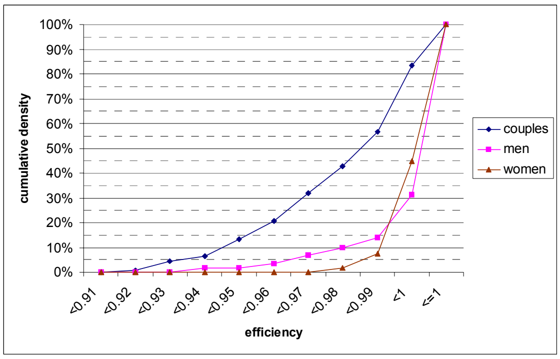

Figure 1: Unitary model singles and couples: cumulative distribution function of improved violation index

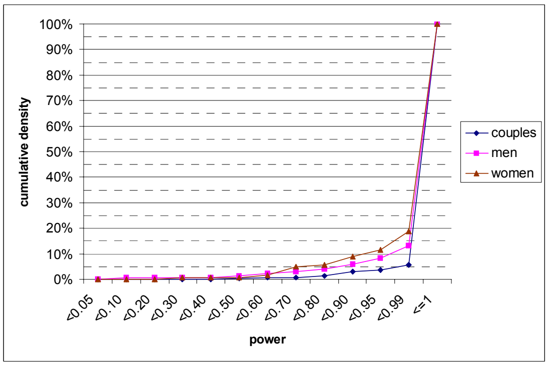

Figure 2: Unitary model singles and couples: cumulative distribution function of power

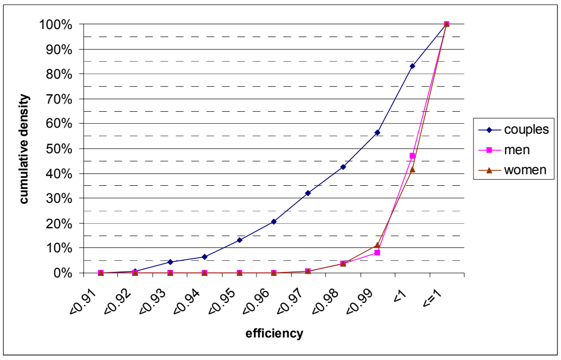

Figure 3: Unitary versus collective model couples: cumulative distribution function of improved violation index

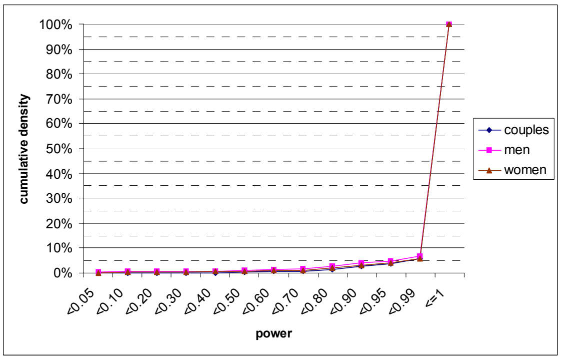

Figure 4: Unitary versus collective model couples: cumulative distribution function of power

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.

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

	Imp. viol. index	**Power index**
women vs. couples	0.000	0.991
men vs. couples	0.000	0.958

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.

Table 3: Sample statistics

	Couples		Single males		Single females	
Variable	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Weekly working hours male	37.90	7.20	37.27	4.73		
Weekly working hours female	34.62	7.62			35.17	7.51
Net hourly wage rate male	9.26	3.45	8.87	4.04		
Net hourly wage rate female	7.23	2.51			8.14	3.17
Age male	35.82	8.67	37.56	8.34		
Age female	33.79	8.45			37.48	9.91
Dummy primary education male	0.04	0.20	0.09	0.28		
Dummy secondary education male	0.54	0.50	0.58	0.49		
Dummy non-academic higher education male	0.30	0.46	0.26	0.44		
Dummy academic higher education male	0.12	0.32	0.07	0.26		
Dummy primary education female	0.05	0.22			0.09	0.29
Dummy secondary education female	0.52	0.50			0.41	0.49
Dummy non-academic higher education female	0.33	0.47			0.34	0.48
Dummy academic higher education female	0.09	0.28			0.16	0.37
Dummy Brussels Capital Region	0.05	0.22	0.10	0.30	0.20	0.40
Dummy Flanders Region	0.73	0.44	0.59	0.49	0.51	0.50
Dummy Walloon Region	0.21	0.41	0.31	0.46	0.29	0.46
Weekly nonlabour income	19.45	54.33	13.54	46.36	16.36	46.75
Dummy wave 1997	0.51	0.50	0.61	0.49	0.51	0.50

Note: Source: Belgian Socio-Economic Panel 1992-1997. Monetary variables are in 1997 euro.