KU LEUVEN

DEPARTMENT OF ECONOMICS

Are consumers rational? Shifting the burden of proof

Laurens CHERCHYE, Thomas DEMUYNCK, Bram DE ROCK and Joshua LANIER

FACULTY OF ECONOMICS AND BUSINESS

DISCUSSION PAPER SERIES DPS20.09

JUNE 2020

Are consumers rational? Shifting the burden of proof

Laurens Cherchye* Thomas Demuynck† Bram De Rock‡
Joshua Lanier§

June 5, 2020

Abstract

We present a statistical test for the hypothesis of rational utility maximization on the basis of nonparametric revealed preference conditions. Our test is conservative for the utility maximization hypothesis. We take as null hypothesis that the consumer behaves randomly, and as alternative hypothesis that she is approximately utility maximizing. Our statistical test uses a permutation method to operationalize the principle of random consumer behavior. We show that the test has an asymptotic power of one against the alternative hypothesis of approximately utility maximizing behavior. We also provide simulated power results and two empirical applications (to experimental and observational data, respectively).
Keywords: utility maximization, revealed preferences, random behavior, permutation test.

1 Introduction

Do consumers act as rational utility maximizers? Despite the huge surge in behavioral economics, the assumption of utility maximization remains a cornerstone of most models

*Department of economics, University of Leuven. E. Sabbelaan 53, B-8500 Kortrijk, Belgium. E-mail: laurens.cherchye@kuleuven.be. Laurens Cherchye gratefully acknowledges the Fund of Scientific Research Flanders (FWO-Vlaanderen) and the Research Fund of the KU Leuven for financial support.

†ECARES, Université Libre de Bruxelles. Avenue F. D. Roosevelt 50, CP 114, B-1050 Brussels, Belgium. E-mail: thomas.demuynck@ulb.be. Thomas Demuynck acknowledges financial support by the Fonds de la Recherche Scientifique-FNRS under grant nr F.4516.18 and EOS project 30544469.

‡ECARES, Université Libre de Bruxelles, and Department of Economics, University of Leuven (KU Leuven). Avenue F. D. Roosevelt 50, CP 114, B-1050 Brussels, Belgium. E-mail: bram.de.rock@ulb.be. Bram De Rock gratefully acknowledges FWO and FNRS for their financial support.

§ECARES, Université Libre de Bruxelles. Avenue F. D. Roosevelt 50, CP 114, B-1050 Brussels, Belgium. E-mail: joshua.lanier@ulb.ac.be.

in economics. Given its importance, it is crucial to check whether actual consumer behavior is at least close to rationality. Revealed preference theory provides an attractive framework to do so. In his seminal contribution, Afriat (1967) showed that a finite data set on observed prices and consumed bundles is rationalizable by the model of utility maximization if and only if it satisfies GARP (Generalized Axiom of Revealed Preference).[1] A most attractive conceptual feature of the revealed preference approach is that it is intrinsically nonparametric, meaning that it abstains from imposing any, typically nonverifiable, functional structure on the consumer's utility function. From a practical perspective, it has the additional advantage that it can be meaningfully applied even to small data sets. For example, GARP can be rejected with only two observed consumption bundles. These two features are why revealed preference methods are frequently used for testing the hypothesis of utility maximizing consumption behavior.

In applications, revealed preference tests usually start from a finite set of observed consumption decisions (prices and quantities) for a given individual, and then verify whether these observations satisfy some combinatorial condition (like GARP). The result of these deterministic tests is either a 'yes' or a 'no'. A 'yes' means that there exists a utility function that *exactly* rationalizes all observed consumption choices as utility maximizing, while a 'no' indicates the opposite. However, as argued by Varian (1991), exact utility maximization might not be a very interesting hypothesis. What we really want to know is whether consumers exhibit *nearly* optimizing behavior, meaning that the rationality hypothesis provides a useful approximation of their observed behavior. As a response to the sharp nature of the deterministic revealed preference tests, it is nowadays customary to complement the tests with a goodness-of-fit measure that quantifies how close the observed behavior is to passing the strict revealed preference conditions. The most popular measure in the applied literature is Afriat's Critical Cost Efficiency Index (CCEI). This CCEI takes values between 0 and 1, with higher values indicating that behavior is closer to exact utility maximization (see Section 2 for a formal definition). Intuitively, one minus the CCEI equals the fraction that the consumer is allowed to waste in each observed consumption decision while still being labeled as approximately utility maximizing.[2]

[1] To be precise, Afriat (1967) originally derived the empirical equivalence between utility maximization and a "cyclical consistency" condition. Varian (1982) has shown the equivalence between GARP and Afriat's cyclical consistency condition. Afriat (1967) built on earlier work of Samuelson (1938) and Houthakker (1950). See also Diewert (1973) and Varian (1982) for detailed and insightful discussions of Afriat's pioneering article, and Chambers and Echenique (2016) for a recent review of the literature.

[2] See Choi, Kariv, Müller, and Silverman (2014) and Dziewulski (2019) for more discussion.

Our contribution. Despite the popularity of the CCEI in applied work, there does not exist a method that determines the CCEI values for which we can reasonably conclude that the model of (approximate) utility maximization provides a good description of the observed behavior. The current paper aims to fill this gap, by providing a statistical test of individual utility maximization.[3] More specifically, we propose to use the CCEI as a statistic for testing the null hypothesis of irrational, random consumption behavior against the alternative hypothesis of approximate utility maximization.[4] Our testing method calculates critical CCEI values to determine the statistical support for the rationality hypothesis.

Our method shifts the burden of proof for the utility maximization hypothesis: we only reject irrational/random consumer behavior if there is substantially strong evidence favoring approximate utility maximization. Our default hypothesis is not that the consumer is utility maximizing but, instead, that she is irrational. To be more precise, the null hypothesis of our test specifies that the consumer's purchasing decisions cannot be distinguished from random behavior. As we motivate in more detail in Section 3, we model irrational behavior by assuming that the consumer randomly draws consumption rays from some distribution that is independent from the budget (i.e. prices and income).[5] Our alternative hypothesis is that the consumer is approximately utility maximizing (as characterized by a specific CCEI value). This means that our framework is conservative for the utility maximization hypothesis. The underlying argument is that, if a data set cannot be distinguished from random behavior, then it should not be treated as arising from the process of utility maximization.

Our testing procedure relies on a permutation approach to operationalize the principle of irrational, random choice behavior.[6] The idea of the test is fairly straightforward. For a given data set on prices and quantities, we consider the population of data sets

[3]Existing studies have developed statistical tests of utility maximization for populations of individuals. See, for example, the recent paper of Kitamira and Stoye (2018). In this study, we focus on individuals rather than populations of individuals.

[4]We focus on the CCEI as our test statistic as this measure is well-known and easily computable. Importantly, however, the use of our testing method is not restricted to the CCEI. One may equally well use other goodness-of-fit measures that have been proposed in the revealed preference literature. Examples include the Houtman-Maks index (Houtman and Maks, 1985), the Varian index (Varian, 1991), the money pump index (Echenique, Lee, and Shum, 2011), the swaps index (Apesteguia and Ballester, 2015) or the minimum cost index (Dean and Martin, 2016).

[5]For a given consumption bundle $(q_1, \ldots, q_L)$ containing L goods, the consumption ray equals the vector $(r_1, \ldots, r_L)$ where $r_i = \frac{q_i}{\sum_{j=1}^{L} q_j}$. Plotting all consumption bundles with the same ray vector $(r_1, \ldots, r_L)$ obtains a line (ray) through the origin that passes through the bundle $(q_1, \ldots, q_L)$. Importantly, our procedure is readily adapted to apply to alternative models of irrational/random behavior (for example, drawing random budget shares). In this respect, we refer to our discussion on specifying the null hypothesis of our statistical test in Section 3.

[6]See, for example, Pesarin and Salmaso (2010) for a review of the permutation testing approach.

that is obtained by fixing the budgets but permuting the consumption rays over the different observations. If the consumer is really randomizing, then the CCEI of the observed data set is equally likely to be realized as any CCEI of these permuted data sets. As such, the distribution of the CCEIs over the permuted data sets provides the distribution for the CCEI of the true data set under the null hypothesis, conditional on the realized observations of prices and quantities.

As our test belongs to the family of permutation tests, it has the specific advantage that it is exact for any sample size. This is particularly convenient in the current setting, as individual revealed preference tests are usually conducted for a small number of observations. For example, our own empirical exercises consider real-life panel data with 26 waves per subject and experimental data with 50 choice observations per subject. In this regard, we also establish a theoretical lower bound on the power of our permutation test with respect to the alternative hypothesis that the observed behavior is approximately utility maximizing. This lower bound converges to one as the number of observations increases.

Relation to power and predictive success. Our approach shares some resemblance with Bronars' (1987) procedure for measuring the power of revealed preference tests. Similar to our procedure, Bronars' power index starts from the idea that irrational behavior can be modeled as random behavior.[7] Computing this index starts by generating a large number of random data sets, and the index is calculated as the fraction of these random data sets that violate (approximate) utility maximization. In the operationalization of Bronars' procedure, random behavior is usually simulated by drawing consumption bundles at random from the budget hyperplane. This, however, implies an ad hoc reliance on some distribution to simulate random behavior, and different distributions may generate different power results. In addition, the chosen distribution may bear little resemblance to the actual distribution of consumption, even if the consumer is truly drawing consumption bundles at random. By contrast, our notion of irrational behavior allows subjects to draw consumption rays at random from any distribution.

From this perspective, our permutation method is more closely related to the 'bootstrap' method that has also been used for measuring the power of revealed preference tests (see, for example, Andreoni and Miller (2002)). Although this bootstrapping approach does away with the reliance on some arbitrary distribution, it has –to our knowledge– no theoretical grounding. Another main difference with our procedure is that these power measures are designed to produce an index of the strictness of deterministic revealed preference tests. Lower index values then indicate that the collection

[7]This idea goes back to Becker (1962).

of observed budget sets only allows for rather weak revealed preference tests and, therefore, does not allow for strong statements favoring the rationality hypothesis. Unlike our method, however, the Bronars or bootstrap index cannot be used directly to test whether or not a particular consumer is a utility maximizer.

Another popular measure in empirical revealed preference analysis is Beatty and Crawford (2011)'s predictive success measure, which is based on an original idea of Selten (1991). This measure is computed as the difference between the pass rate of a revealed preference test (from a population of data sets) and one minus Bronars' power index of this test. Predictive success values close to zero imply that the pass rate for the observed data sets is close to the pass rates for the population of randomly generated data sets. By contrast, values close to one point out that (almost) all observed data pass the revealed preference tests, while the opposite holds for random data. Finally, values below zero indicate that random behavior performs better than actual behavior on the revealed preference tests. Summarizing, Beatty and Crawford's predictive success measure tells us how well a revealed preference test can distinguish between actual behavior and random behavior. However, it remains silent about whether a particular individual behaves according to the utility maximization model or what critical values are to be used to reach that conclusion. This is exactly the distinguishing feature of our procedure. In this sense, we see the two procedures as complementary, each one highlighting a different aspect of the data.

Outline. The remainder of this paper unfolds as follows. Section 2 sets the ground by introducing some basic concepts, and by giving an example that motivates our test. Section 3 formally presents our statistical testing procedure. Section 4 derives a theoretical lower bound on the power of our statistical test. Section 5 discusses simulated power results and provides two empirical applications (on experimental and observational data, respectively). Section 6 contains our conclusion. All our proofs are in the Appendix.

2 Basic concepts

We start by briefly introducing some necessary concepts and notation. Throughout, we will consider a consumption setting with L goods. A revealed preference analysis usually departs from a finite set of observed prices $p^t = [p_1^t, \ldots, p_L^t] \in \mathbb{R}^L_{++}$ and associated quantities $q^t = [q_1^t, \ldots, q_L^t] \in \mathbb{R}^L_+$. The idea is that, at each observation t, the consumer purchased the bundle q^t under the prevailing prices p^t. A data set is denoted by $D = (q^t, p^t)_{t \leq T}$.

GARP and CCEI. We say that the bundle q^t is revealed preferred to the bundle q^v if $p^t \cdot q^t \geq p^t \cdot q^v$. We denote this by $q^t \, R \, q^v$. In words, the bundle q^t was chosen at observation t while q^v was also attainable (for the given expenditure $p^t \cdot q^t$ and prices p^t). Similarly, a bundle q^t is strictly revealed preferred to q^v if $p^t \cdot q^t > p^t \cdot q^v$, which we denote by $q^t \, P \, q^v$. Intuitively, q^t was chosen although q^v was equally affordable together with some additional money left for the consumer.

A data set D satisfies the Generalized Axiom of Revealed Preference (GARP) if there is no 'strict' cycle in the revealed preference relation: for any sequence of observations $t_1, \ldots, t_M \leq T$,

$$q^{t_1} \, R \, q^{t_2} \, R \, \ldots \, R \, q^{t_M} \text{ implies not } q^{t_M} \, P \, q^{t_1}.$$

Afriat (1967) has shown that the observed behavior (captured by the data set D) can be rationalized as maximizing a well-behaved (i.e. increasing, continuous and quasi-concave) utility function if and only if the set D satisfies GARP.

If a data set does not satisfy GARP, we may consider a weakening of the sharp GARP condition. As indicated in the Introduction, a popular way to do so makes use of the Critical Cost Efficiency Index (CCEI). To formally define this CCEI, we consider the relations $q^t \, R^e \, q^v$ if $e(p^t \cdot q^t) \geq p^t \cdot q^v$ and $q^t \, P^e \, q^v$ if $e(p^t \cdot q^t) > p^t \cdot q^v$, which make use of a prespecified 'efficiency' value $e \in [0, 1]$. Intuitively, the revealed preference relations R^e and P^e imply a weakening of the relations R and P, as q^t is now said to be (strictly) revealed preferred to q^v only if q^v was available when the budget at observation t was decreased by a fraction $(1 - e)$. We say that a data set D satisfies e-GARP if, for all sequences of observations $t_1, \ldots, t_M \leq T$,

$$q^{t_1} \, R^e \, q^{t_2} \, R^e \, \ldots \, R^e \, q^{t_M} \text{ implies not } q^{t_M} \, P^e \, q^{t_1}.$$

Obviously, for $e = 1$ we have that e-GARP coincides with GARP. Moreover, any data set satisfies e-GARP for $e = 0$. More generally, if a data set D satisfies e-GARP, then it will satisfy e'-GARP for any efficiency value $e' \leq e$. This calls for defining the highest value of e such that a data set still satisfies e-GARP. This value gives us the CCEI, which we denote by $\tau(D)$ for a data set D:

$$\tau(D) = \sup\{e \in [0, 1] : D \text{ satisfies } e\text{-GARP}\}.$$

Varian (1990) proposed the CCEI as a goodness-of-fit measure in empirical revealed preference analysis. The higher the value of the CCEI, the closer the observed data set is to satisfying GARP. As indicated in the Introduction, one minus the CCEI equals the fraction that the consumer is allowed to waste in each observed consumption decision,

while still being labeled as approximately utility maximizing.

Critical CCEI value. A natural question is whether an observed CCEI value is sufficiently high to conclude that the decision maker is approximately utility maximizing and not just picking consumption bundles at random. In the literature, there is no consensus on what value the CCEI should minimally attain to conclude that behavior is (approximately) rational. Varian (1991) mentions the critical value of 0.95, but admits that this is mainly out of sentimental reasons. Choi, Fisman, Gale, and Kariv (2007) use 0.90 based on their results for Bronars' power procedure. Particularly, for their application these authors find that the CCEI is below 0.90 for most randomly generated data sets (using the uniform distribution to simulate random behavior). Most of the other papers in the literature tend to use cut-offs of 0.90, 0.95 or 0.99 (see for example Polisson, Quah, and Renou (2020)).

In the current paper, we set out a framework to define an individual-specific cut-off that is determined as the critical value of a statistical test. To determine this cut-off, we consider a data set that is obtained by random choice. Random choice is modeled by fixing the various budgets but permuting the consumption rays over the different observations. Figure 1 provides an illustration for three observations and two goods. The budgets are given by the solid lines and the bundles are represented by squares. Assume that we observe a consumer who picks the consumption bundles according to the top left panel of the figure. The observed data set then produces three consumption rays depicted by the dashed lines through the origin. If the consumer were irrational and picked her consumption rays from some random distribution, the observed consumption pattern would be equally likely as any consumption pattern in the other 5 panels of Figure 1, which are obtained by permuting the three observed consumption rays over the observed budgets.

Let us denote a permuted data set by D_σ (see Section 3 for a formal definition), and the corresponding CCEI value by $\tau(D_\sigma)$. If the individual chose her consumption bundles by randomly picking rays from some distribution then, conditional on the observed rays, the probability of observing the data set D must have the same likelihood as observing the data set D_σ. This is the main idea behind our permutation test. To put this into practice, we compute the CCEI values for all possible data sets that are obtained by permuting the consumption rays over the observations. If actual behavior picked consumption rays at random, then the CCEI value for the true data set would be a random draw from the distribution of all these CCEI values. Thus, we can reject the hypothesis of random behavior at the significance level α if at most a fraction α of all the permuted data sets have a CCEI value above or equal to $\tau(D)$.

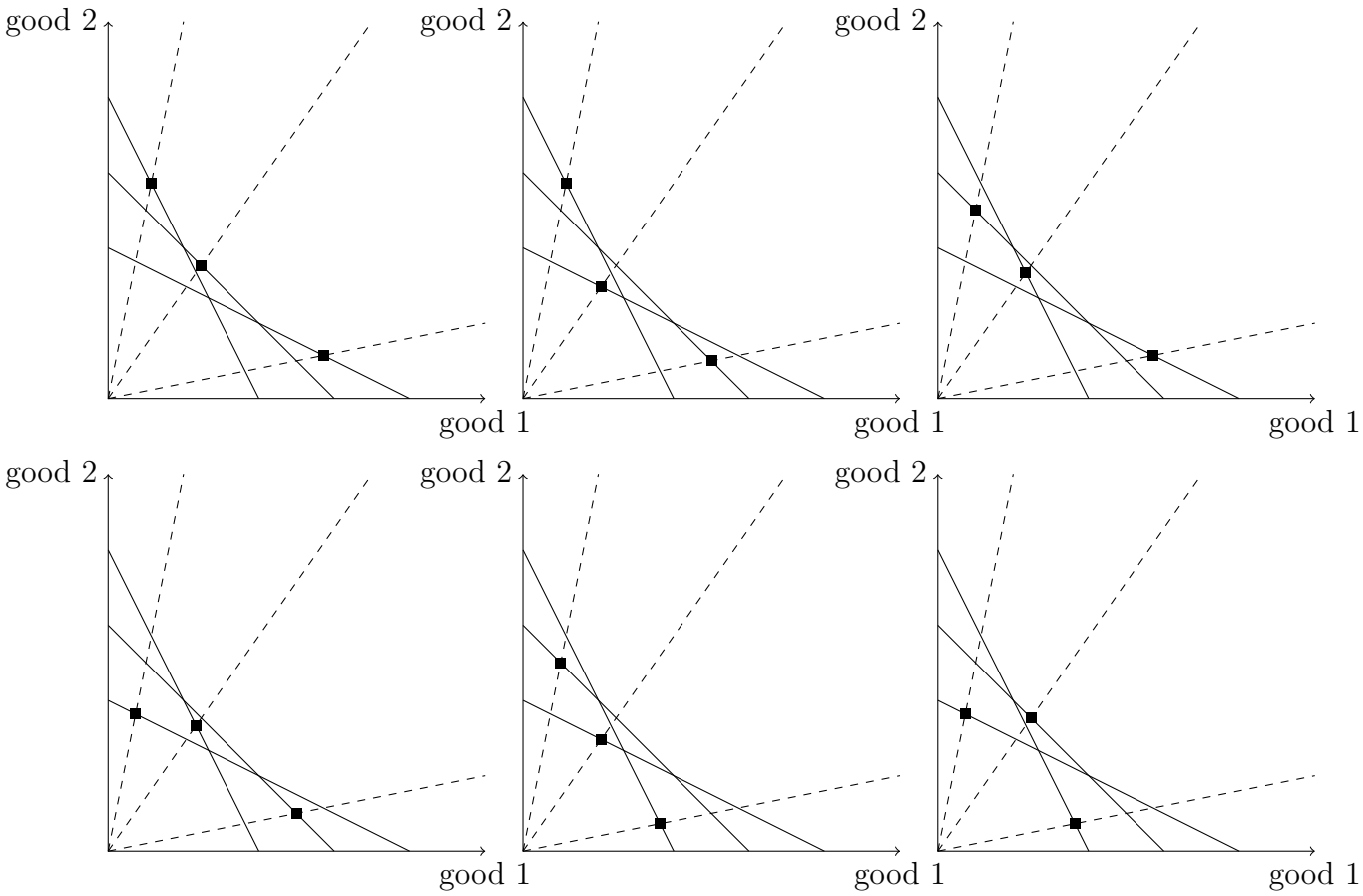

Figure 1: Permuting consumption rays

Example. Figure 2 shows an artificial data set with nine observations and two goods. This data set violates GARP but only to a small degree. In particular, we have that $\tau(D) = 0.984$, indicating that the CCEI is quite close to one. When computing the CCEI for all 362 880 permuted data sets, we find that 2.49% of these data sets have a CCEI that is at least as high as $\tau(D) = 0.984$. In other words, the p-value for the null hypothesis that the consumer was randomly picking consumption shares is 0.0249. We conclude that the null of random consumer behavior cannot be rejected at a significance level of 1%, while it is rejected at the 5% or 10% level.

3 Permutation test

Like before, we assume L goods, and we consider a data set $D = (q^t, p^t)_{t \leq T}$ that consists of T quantities $q^t \in \mathbb{R}^L_+$ and associated prices $p^t \in \mathbb{R}^L_{++}$. A data set contains a fixed number of observations, and we assume that each set D is drawn from a probability

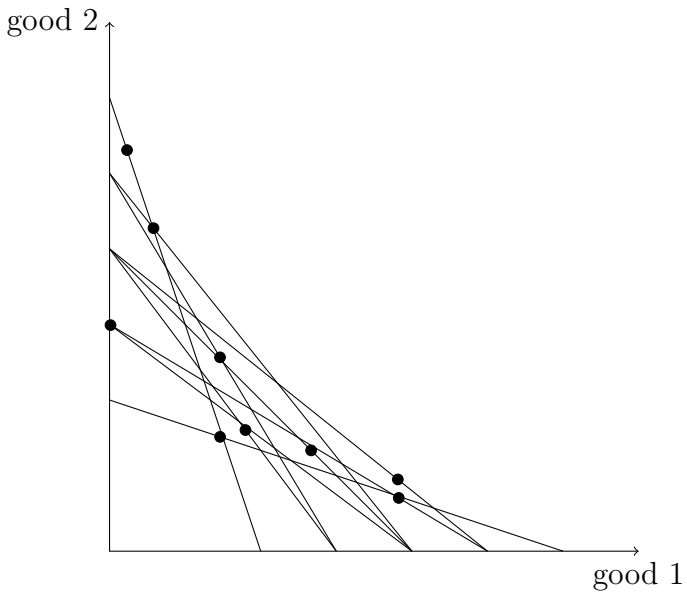

Figure 2: Example

space $(\Omega, \mathcal{B}, P)$, where

$$\Omega = \left\{ (q^t, p^t)_{t \leq T} \in \left(\mathbb{R}^L_+ \times \mathbb{R}^L_{++} \right)^T \right\},$$

is the set of all possible data sets, $\mathcal{B}$ is the Borel sigma-algebra on Ω (i.e. the sigma-algebra generated by the closed subsets of Ω), and P is a probability measure on $(\Omega, \mathcal{B})$. For a given observation t, with quantity-price vector (q^t, p^t), we can define the vector $r^t = [r^t_1, \ldots, r^t_L] \in \mathbb{R}^L_+$ as

$$r^t_i = \frac{q^t_i}{\sum_{j=1}^L q^t_j}.$$

We call r^t the ray of the bundle q^t. Next we define the total expenditure at observation (q^t, p^t) as $m^t = \sum_{i=1}^L p^t_i q^t_i$. Note that there is a one-to-one relationship between the observed vector (q^t, p^t) and the triple (r^t, p^t, m^t).[8] Given this equivalence, we will often use $D = (q^t, p^t)_{t \leq T}$ and $D = (r^t, p^t, m^t)_{t \leq T}$ interchangeably in what follows.

Null hypothesis. As described above, our statistical test concentrates on two types of consumers: (approximate) utility maximizers and random consumers. We take as null hypothesis that the observed data set D was generated by a consumer who randomly draws consumption rays from some distribution.

[8]The bundle q^t can be derived from the ray r^t and the budget m^t by the transformation $q^t_i = r^t_i m^t / (\sum_{j=1}^L p^t_j r^t_j)$.

Definition 1. *A probability measure P on $(\Omega, \mathcal{B})$ is consistent with random consumption rays if, for every random data set $D = (r^t, p^t, m^t)_{t \leq T}$ and any observation t, r^t is drawn from a distribution that is independent of t, independent from all prices $(p^v)_{v \leq T}$ and independent from all income levels $(m^v)_{v \leq T}$. We denote by $\mathcal{P}_0$ the set of all such probability measures.*

Definition 1 specifies that a consumer behaves irrationally when her consumption choices are defined by randomly picking consumption rays from a particular distribution, irrespective of the particular observation or the observed prices and income levels. At this point, we remark that our null hypothesis does not necessarily exclude utility maximizing behavior. For example, a consumer with a Leontief utility function (i.e. perfect complements) will always consume on a fixed ray and, therefore, her choices will coincide with choosing a ray from a distribution with a point mass of one at a single ray vector. Except from such fairly pathological cases, however, our null hypothesis does exclude rational behavior as consumption will usually depend on prices and income levels.

We acknowledge that the random rays null hypothesis that we use represents only one of many possible ways to model random behavior. For example, Becker (1962), Bronars (1987) and Beatty and Crawford (2011) equate random behavior as picking budget shares instead of rays. Our test can easily be altered to allow for a null hypothesis where the consumer selects random budget shares instead of random consumption rays. This random shares hypothesis models a consumer who randomizes over the fractions of her budget she wishes to allocate to each commodity without actually looking at the underlying price. This can reasonably be thought to describe a consumer making choices over aggregated categories of goods. For example, one might imagine a consumer who decides to spend 20 percent of the budget on food, 30 percent on energy, and so forth.

To some extent, the rays versus shares hypothesis appears to relate to the aggregation level of the goods. If we investigate purchase behavior at the micro level (over particular goods) the rays hypothesis seems more natural: in order to make meals, one usually needs to buy ingredients in fixed proportions. The total amount of food produced will then determine the budget and the distribution over meals determines the distribution over rays. On the other hand, when looking at consumption allocations over more aggregated group categories, then the shares hypothesis might be more appropriate. As the goods in our empirical exercises in Section 5 do not involve a substantial amount of aggregation, we prefer to stick to the random rays hypothesis as being more appropriate for these applications.

Alternative hypothesis. As alternative hypothesis we use that the consumer is approximately utility maximizing. As we informally introduced in Section 2, we consider a general type of approximate utility maximization that accounts for the possibility of small inefficiencies in consumer behavior, which is captured by an efficiency level $e \in [0,1]$. Higher e-values indicate that the consumer is closer to utility maximization. More specifically, we use the following definition.

Definition 2. *Let $U : \mathbb{R}^L \to \mathbb{R}$ be a well-behaved (i.e. increasing, continuous and quasi-concave) utility function. Let $E^{U,e} \subseteq \Omega$ be the set of all data sets $D = (q^t, p^t)_{t \leq T} \in \Omega$ such that*

$$U(q^t) \geq \max_{\tilde{q}} U(\tilde{q}) \text{ s.t. } p^t \cdot \tilde{q} \leq e\,(p^t \cdot q^t), \qquad \forall t \leq T. \tag{1}$$

We say that the probability measure P on $(\Omega, \mathcal{B})$ is consistent with e-utility maximization if there exists a well-behaved utility function U such that

$$P\Big(D \in E^{U,e}\Big) = 1.$$

We denote the set of all measures P that are consistent with e-utility maximization by $\mathcal{P}_1^e$.

The set $E^{U,e}$ in this definition contains all data sets D such that the utility $U(q^t)$ received at observation t equals at least the maximally attainable utility after removing a fraction $(1-e)$ from the consumer's budget.[9] Definition 2 states that a probability measure P is consistent with e-utility maximizing behavior if it assigns, for a well-behaved utility function U, a probability of unity to drawing a data set from $E^{U,e}$.

The following result connects the concept of e-utility maximization in Definition 2 to the e-GARP concept (and, thus, the CCEI) that we introduced in Section 2.[10]

Theorem 1. *Let $e \in [0,1]$ and $D = (q^t, p^t)_{t \leq T}$ a data set in Ω. Then, there exists a well-behaved utility function U such that $D \in E^{U,e}$ if and only if D satisfies e-GARP.*

From Theorem 1 we obtain that, if $P \in \mathcal{P}_1^e$, then any data set D drawn from the probability space $(\Omega, \mathcal{B}, P)$ must satisfy e-GARP with probability one and, thus, have a CCEI value that is at least equal to e. Indeed, if $P \in \mathcal{P}_1^e$, then we know that there exists a well-behaved utility function U such that

$$1 = P(D \in E^{U,e}) \leq P(D \text{ satisfies } e\text{-GARP}) \leq P(\tau(D) \geq e),$$

[9] Given continuity of U, one can show that the set $E^{U,e}$ is closed, so it is measurable.

[10] Although Theorem 1 is well-known and fairly intuitive, we are not aware of any formal proof in the literature. Therefore, we have included a proof in Appendix A.1.

where the first inequality follows from Theorem 1 and the second from our definition of the CCEI. Conversely, if $P(\tau(D) \geq e) < 1$, then it must be that $P \notin \mathcal{P}_1^e$.

Hypothesis test. Summarizing, our statistical test considers the following hypothesis:

$H_0 : P \in \mathcal{P}_0$, i.e. the consumer is a random consumer.

$H_1 : P \in \mathcal{P}_1^e$, i.e. the consumer is a e-utility maximizer, for some $e \in [0, 1]$.

We now formally introduce our statistical procedure to test H_0 against H_1. Let σ represent a permutation on $\{1, \ldots, T\}$. Then, we define a permuted data set D_σ as

$$D_\sigma = \left(r^{\sigma(t)}, p^t, m^t\right)_{t \leq T} \equiv \Big(\left(r^{\sigma(1)}, p^1, m^1\right), \ldots, \left(r^{\sigma(T)}, p^T, m^T\right)\Big).$$

Let Π denote the set of all permutations on $\{1, \ldots, T\}$. The total number of permutations in Π equals $T!$. Then, our testing procedure goes as follows.

Procedure 1. *Let $\alpha \in (0, 1)$. Reject $H_0 : P \in \mathcal{P}_0$ in favor of $H_1 : P \in \mathcal{P}_1^e$ at the significance level α if*

$$\phi_\alpha(D) = 1,$$

where

$$\phi_\alpha(D) = \mathbf{1}\left[\frac{\left|\left\{\sigma \in \Pi : \tau(D_\sigma) \geq \tau(\mathcal{D})\right\}\right|}{T!} \leq \alpha\right]$$

and $\mathbf{1}[.]$ is the indicator function that equals 1 if the expression between brackets is true and 0 otherwise.

In words, our testing procedure considers all possible permutations D_σ of the data set D. We then compute the fraction of permuted data sets of which the CCEI (i.e. $\tau(D_\sigma)$) is at least as high as the CCEI of the actual data set (i.e. $\tau(D)$). If this fraction is less than or equal to α, then we reject the null hypothesis at the significance level α.

The next result motivates the theoretical validity of Procedure 1, by showing that the probability of making a Type-1 Error is at most α.[11]

Theorem 2. *Let $\alpha \in (0, 1)$ and $P \in \mathcal{P}_0$. Then,*

$$P(\phi_\alpha(D) = 1) = \mathbb{E}_P\left[\phi_\alpha(D)\right] \leq \alpha.$$

[11] From the continuity of the CCEI measure $\tau(D)$, it follows that ϕ_α is a measurable function.

4 Statistical power

Theorem 2 motivates our permutation test by characterizing its size. However, it may still be that the test has low power, i.e. the probability of rejecting the null hypothesis might be low even if the consumer is actually approximately utility maximizing. To address this issue, we establish a theoretical lower bound on the statistical power of our permutation test. We also specify conditions under which this lower bound converges to one when the sample of observations grows.

Given that our power result relies on large sample statistics, we need to limit ourselves to a somewhat more restrictive data generating process. In particular, we assume that the various observations in a data set are i.i.d. draws from some common distribution.[12] Towards this end, we fix a probability space $(\widehat{\Omega}, \widehat{\mathcal{B}}, \widehat{P})$, where

$$\widehat{\Omega} = \{(q,p) \in \mathbb{R}^L_+ \times \mathbb{R}^L_{++}\},$$

$\widehat{\mathcal{B}}$ is the Borel sigma-algebra on $\widehat{\Omega}$, and $\widehat{P}$ is a probability measure on $(\widehat{\Omega}, \widehat{\mathcal{B}})$. From now on, we assume that a random data set of size T is obtained by taking T independent draws from $(\widehat{\Omega}, \widehat{\mathcal{B}}, \widehat{P})$. In other words, $D = (q^t, p^t)_{t \leq T}$ is a random draw from the product probability space $(\widehat{\Omega}_T, \widehat{\mathcal{B}}_T, \widehat{P}_T)$, where

$$\widehat{\Omega}_T = \underbrace{\widehat{\Omega} \times \widehat{\Omega} \times \ldots \times \widehat{\Omega}}_{T \text{ times}},$$
$$\widehat{\mathcal{B}}_T = \underbrace{\widehat{\mathcal{B}} \otimes \widehat{\mathcal{B}} \otimes \ldots \otimes \widehat{\mathcal{B}}}_{T \text{ times}},$$
$$\widehat{P}_T = \underbrace{\widehat{P} \times \widehat{P} \times \ldots \times \widehat{P}}_{T \text{ times}}.$$

Given a well-behaved utility function U, we let $\widehat{E_T}^{U,e}$ be the set of observations for which

$$U(q^t) \geq \max_{\tilde{q}} U(\tilde{q}) \text{ s.t. } p^t \cdot \tilde{q} \leq e(p^t \cdot q^t), \qquad \forall t \leq T,$$

and we can define $\widehat{\mathcal{P}}^e_{T,1}$ to be the set of all probability distributions $\widehat{P}_T$ on $(\widehat{\Omega}_T, \widehat{\mathcal{B}}_T)$ such that

$$\widehat{P}_T\Big((q,p)_{t \leq T} \in \widehat{E_T}^{U,e}\Big) = 1,$$

for some well-behaved utility function U. Our alternative hypothesis is that $\widehat{P}_T \in \widehat{\mathcal{P}}^e_{T,1}$

[12] This i.i.d. assumption may seem restrictive, especially for real-life data, where it is known that prices exhibit some persistence. On the other hand, for experimental settings where budgets are generated at random, this i.i.d. assumption is quite natural.

for some $e \in [0,1]$.

Now let us fix an efficiency level $e \in [0,1]$ and consider a probability measure $\widehat{P}_T$ in agreement with the alternative hypothesis, i.e. $\widehat{P}_T \in \widehat{\mathcal{P}}^e_{T,1}$. Further, we let $D = \big((r^1,p^1,m^1),(r^2,p^2,m^2)\big)$ be a data set of size 2 that is drawn randomly from $(\widehat{\Omega}_2, \widehat{\mathcal{B}}_2, \widehat{P}_2)$. Then, consider the permuted data set

$$\widetilde{D} = \big((r^2,p^1,m^1),(r^1,p^2,m^2)\big).$$

As before, it directly follows that

$$1 = \widehat{P}_2\big(D \in \widehat{E}^{U,e} \times \widehat{E}^{U,e}\big) \leq \widehat{P}_2\big(D \text{ satisfies } e\text{-GARP}\big) \leq \widehat{P}_2(\tau(D) \geq e).$$

On the other hand, as $\widetilde{D}$ was obtained by permuting the original consumption rays, it may well be that the event $\tau(\tilde{D}) \geq e$ has a probability below one. Let us define

$$\widetilde{\pi}_e = \widehat{P}_2\big(\tau(\widetilde{D}) \geq e\big).$$

Next, let $D = \big((r^1,p^1,m^1),(r^2,p^2,m^2),(r^3,p^3,m^3)\big)$ be a data set of size 3 that is drawn randomly from $(\widehat{\Omega}_3, \widehat{\mathcal{B}}_3, \widehat{P}_3)$, and consider the permuted data set $\overline{D}$ of size 2 that is obtained from D as

$$\overline{D} = \big((r^2,p^1,m^1),(r^3,p^2,m^2)\big).$$

Similar to before, we define

$$\overline{\pi}_e = \widehat{P}_3\big(\tau(\overline{D}) \geq e\big).$$

Again, it is quite likely that $\overline{\pi}_e < 1$. The next result shows that, if both $\widetilde{\pi}_e$ and $\overline{\pi}_e$ are below one, then the asymptotic power of our permutation test for the alternative hypothesis $\widehat{P} \in \widehat{\mathcal{P}}^e_1$ equals unity. More generally, the result defines a theoretical lower bound on the power of our test.

Theorem 3. *Let $\alpha \in (0,1)$ and $e \in [0,1]$. Further, let D be a data set of size T that is drawn randomly from $(\widehat{\Omega}_T, \widehat{\mathcal{B}}_T, \widehat{P}_T)$. Assume that $\widehat{P}_T \in \widehat{\mathcal{P}}^e_{T,1}$. Then, the probability of not rejecting the null hypothesis under $\widehat{P}_T$ is bounded by*

$$\widehat{P}_T\big(\phi_\alpha(D) = 0\big) = \mathbb{E}_{\widehat{P}_T}\big(1 - \phi_\alpha(D)\big) \leq \frac{4}{\alpha}(\pi_e)^{\frac{T}{4}},$$

where

$$\pi_e = \max(\widetilde{\pi}_e, \overline{\pi}_e, 1/2).$$

5 Empirical exercises

Theorem 3 is a large sample result. In what follows, we first conduct a simulation exercise that investigates the power of our permutation test in the finite sample case. Subsequently we demonstrate the empirical usefulness of our testing procedure by applying it to the experimental data set of Fisman, Kariv, and Markovits (2007) and the real-life Stanford Basket data set that was also used by Echenique, Lee, and Shum (2011). To illustrate the versatility of our approach, we end by exploring the added value of considering the more restricted class of quasi-linear preferences.

The permutation test that we outlined in Section 3 starts by calculating the CCEI $\tau(D)$ for a data set D of observed prices and quantities associated with a single consumer. Subsequently, it permutes the budget rays across observations, and computes the CCEI $\tau(D_\sigma)$ for each permuted data set D_σ. This means that we must compute $\tau(D_\sigma)$ for each of the $(T!)$ possible permutations σ. However, for large enough data sets, this quickly becomes computationally intractable. For example, for a data set of 50 observations, this requires $(50!) \geq 3 \times 10^{64}$ permutations. To ensure computational feasibility, it is standard practice in the literature on permutation tests to take a large enough sample of random permutations when the number of observations becomes too large. In our following exercises, our test uses all possible permutations when the number of observations equals at most 7. In the other cases, we randomly sample 10 000 permutations with replacement. In order to speed up our computations, we further employ the following 'heuristic': if after running the test using 1 000 permutations we find a p-value strictly greater than 0.2, we abort the test and report the results using only these 1 000 permutations. This saves us the trouble of refining our p-value when there is little chance of ever approaching the 10% significance level.

The CCEI-value of a data set is computed using a standard binary search algorithm. We choose the number of iterations that guarantees that the CCEI is calculated to within an error of 2^{-17} of the true value. We make sure to always test the data set for GARP so that if the data set is perfectly rationalizable, we return a CCEI value of 1.

Simulated data. To compute the power of our test, we need to generate data that are consistent with e-GARP for chosen values of e. To generate a budget set when there are L goods we start by drawing L numbers uniformly from the interval $[1, 10]$.

	$e = .99$				$e = .95$				$e = .90$			
		α				α				α		
	T	0.10	0.05	0.01	T	0.10	0.05	0.01	T	0.10	0.05	0.01
$L = 2$	6	0.12	0.02	0.00	$L = 2$	6	0.07	0.02	0.00	$L = 2$	6	0.03	0.00	0.00
	8	0.55	0.22	0.00		8	0.34	0.08	0.00		8	0.25	0.07	0.00
	10	0.88	0.63	0.14		10	0.77	0.44	0.03		10	0.63	0.39	0.03
	14	1.00	1.00	0.81		14	0.98	0.90	0.53		14	0.92	0.75	0.24
	20	1.00	1.00	1.00		20	1.00	1.00	0.96		20	1.00	0.99	0.80
$L = 4$	6	0.25	0.06	0.00	$L = 4$	6	0.25	0.07	0.00	$L = 4$	6	0.22	0.04	0.00
	8	0.93	0.61	0.06		8	0.81	0.49	0.01		8	0.68	0.34	0.02
	10	1.00	0.96	0.43		10	0.96	0.89	0.34		10	0.86	0.68	0.16
	14	1.00	1.00	1.00		14	1.00	1.00	0.96		14	1.00	1.00	0.77
	20	1.00	1.00	1.00		20	1.00	1.00	1.00		20	1.00	1.00	0.99
$L = 8$	6	0.58	0.25	0.01	$L = 8$	6	0.53	0.22	0.01	$L = 8$	6	0.40	0.15	0.01
	8	0.99	0.87	0.18		8	0.97	0.82	0.18		8	0.89	0.75	0.09
	10	1.00	1.00	0.88		10	0.99	0.99	0.76		10	0.98	0.90	0.60
	14	1.00	1.00	1.00		14	1.00	1.00	1.00		14	1.00	1.00	0.95
	20	1.00	1.00	1.00		20	1.00	1.00	1.00		20	1.00	1.00	1.00
$L = 16$	6	0.79	0.53	0.08	$L = 16$	6	0.82	0.45	0.12	$L = 16$	6	0.82	0.48	0.05
	8	1.00	0.98	0.50		8	1.00	0.95	0.45		8	1.00	0.95	0.53
	10	1.00	1.00	0.97		10	1.00	1.00	0.95		10	1.00	1.00	0.95
	14	1.00	1.00	1.00		14	1.00	1.00	1.00		14	1.00	1.00	1.00
	20	1.00	1.00	1.00		20	1.00	1.00	1.00		20	1.00	1.00	1.00

Table 1: Power for simulated data

Denote these numbers by $\alpha_1, \ldots, \alpha_L$. Next, for each good i we set the price to $10/\alpha_i$, and we set the expenditure level for the budget equal to 10. This ensures that if the consumer allocates her income to good i, she could purchase exactly α_i units. Once the budgets are selected, we generate 100 random data sets satisfying e-GARP by using the algorithm set out in Appendix B.

Table 1 contains the simulation results for various numbers of goods ($L = 2, 4, 8, 16$) and different number of observations ($T = 6, 8, 10, 14, 20$). The different cells reveal the power of our statistical test for alternative combinations of L, T, e and α. For example, the cell ($T = 8, L = 2$ with $e = 0.99$ and $\alpha = 0.10$) has a value 0.55. This says that for 55 percent of our simulations, we reject the null hypothesis at the 10%-significance level. Generally, we find that the power of our permutation test increases in the number of goods (L) and the number of observations (T). The power is close to 1 as soon as we have 20 observations. We conclude that our test has sufficient power whenever T is moderately large. This especially holds true when the number of goods L also gets large.

Experimental data. An advantage of experimental data is that they allow for gathering a high number of consumption observations for one and the same individual at low cost. In addition, the experimental designer has full control over the various budgets faced by the experimental subjects. This type of data is exactly in line with the setting in our simulation exercise above, which motivates that our procedure has sufficient power.

We illustrate this for the data set on giving versus keeping of Fisman, Kariv, and Markovits (2007). This experiment was designed to investigate individual preferences

for giving by exposing subjects to a series of dictator games under varying incomes and conversion rates between giving and keeping.[13] The data set has two components. The first component contains information for 76 subjects (i.e. 76 consumers) on 50 choices between keeping and giving to one other individual (i.e. 2 goods), and the second component contains information for 65 subjects on choices between keeping and giving to either individual A or individual B (i.e. 3 goods). We refer to Fisman, Kariv, and Markovits (2007) for more details on the data.

Table 2 summarizes our results. Attractively, we find that the experimental data allow us to statistically discriminate between utility maximizing and random behavior. All rejection rates are well above the nominal significance levels, even without imposing specific additional structure on the consumers' utility functions (see also below). For instance, we reject the null hypothesis of random choice behavior at the 0.01 significance level for 72% of the subjects (for the choices with 2 goods) and 83% of the subjects (for the choices with 3 goods). In our opinion, this convincingly demonstrates that our permutation test can have substantial empirical bite in practice.

Real-life data. We next study the Stanford Basket data set that was also used by Echenique, Lee, and Shum (2011). This data set captures consumer expenditures on 14 types of goods that fall in the "food" category, covering the period from June 1991 to June 1993 (i.e. 104 weeks). There are 494 consumers and, after aggregating up to brand level and dropping goods which have no price data for some weeks, we retain a total of 430 goods. We aggregate the data so that one period represents 4 weeks, resulting in a maximum of 26 periods per participant. All our aggregation steps follow Echenique, Lee, and Shum (2011).

If we compute CCEI values for the 494 consumers in the Stanford Basket Data set, we find that 416 (84.2%) have a CCEI value below unity, i.e. they violate the sharp GARP condition. Still, we find that the CCEI values are generally high. The average CCEI equals 0.9504, with a standard deviation of 0.0578. Although the minimum CCEI value equals no more than 0.4278, we observe that the first quartile, median and third quartile amount to 0.93, 0.97 and 0.99. This may suggest that the observed behavior is generally close to approximate utility maximization.

Our test procedure allows us to investigate the statistical support for this claim. In particular, we can use our procedure to assess for which subjects we reject the null of random behavior. The results of this exercise are also given in Table 2. Generally, we

[13] In particular, subjects made several choices by filling in questions of the form: *"Divide X tokens: Hold _____ at a points, and Pass _____ at b points (the Hold and Pass amounts must sum to X)"*. The parameters X, a and b were varied across the decision problems.

find that the statistical support for utility maximizing behavior is rather weak when using a significance level of 1%, with a rejection rate of only 10%. The picture is somewhat more nuanced for the 10% significance level, with a rejection rate of 40%.

Quasi-linear preferences. One possible conclusion from these results in Table 2 is that the restrictions imposed by nearly utility maximizing behavior are often not sufficiently restrictive to significantly distinguish such behavior from purely random behavior. So many types of behavior can count as approximate utility maximization that it is often hard to differentiate it from randomness. To explore this in more detail, we applied our testing procedure when using the (stronger) alternative hypothesis of approximate utility maximization with quasi-linear preferences. Particularly, we say that a utility function U is quasi-linear if there exists an outside good y such that we can write

$$U(q, y) = V(q) + y.$$

The model of quasi-linear utility maximization is substantially more restrictive than the standard utility maximization model. As such, if people effectively behave like approximate quasi-linear utility maximizers, we should more easily detect this when using an appropriate statistical test. For this exercise, we make use of the revealed preference characterization of quasi-linear utility maximization that was developed by Brown and Calsimiglia (2007, Theorem 2.2), which we adapt to our particular setting.[14]

As expected the goodness-of-fit of this more restricted model, measured once more by the CCEI, is significantly lower. In this case, the first quartile, median and third quartile amount to respectively 0.7875, 0.8390 and 0.8810. Next, the results of our statistical test are again summarized in Table 2. It is interesting to note that there are many people for which we reject the null in favor of the alternative hypothesis of nearly quasi-linear utility maximization, but not in favor of the standard nearly utility maximization model, particularly when using a significance level of 1%. This shows that it might often be useful to focus on a more restricted class of utility functions to verify the utility maximization hypothesis. If the observed behavior is consistent with a more restrictive utility maximization model, it will generally be easier to distinguish such optimizing behavior from purely random behavior.

[14]Specifically, the CCEI for quasi-linear utility maximization can be calculated by testing the data for cyclical monotonicity (as defined in Brown and Calsimiglia (2007)), which is the approach we take here.

	Experimental data		Real-life data	
Sign. level	Rejection rates Gen. pref., 2 goods	Rejection rates Gen. pref., 3 goods	Rejection rates Gen. pref.	Rejection rates Quasi-linear pref.
$\alpha = 0.10$	0.88	0.94	0.40	0.49
$\alpha = 0.05$	0.82	0.92	0.30	0.39
$\alpha = 0.01$	0.72	0.83	0.10	0.24

Table 2: Rejection rates for experimental and real-life data

6 Conclusion

We present a novel statistical testing procedure for the hypothesis of (approximately) utility maximization on the basis of nonparametric revealed preference conditions. It allows us to compute critical values for the CCEI for which we cannot reject the hypothesis of rationality for the observed data. A specific feature of our test procedure is that it shifts the burden of proof: we only reject random consumption behavior if there is substantially strong evidence favoring utility maximizing behavior. We take as null hypothesis that consumers behave randomly, and as alternative hypothesis that consumers are approximate utility maximizers. Our statistical test makes use of a permutation method to operationalize the principle of randomization. This permutation procedure is also valid for small samples and allows us to characterize a theoretical lower bound on the power of the test.

We illustrate the practical usefulness of our test for both experimental and observational scanner data. Our application to experimental data shows the use of experiments to statistically discriminate between utility maximizing and random behavior. A main advantage of experimental data is that it allows for gathering a high number of consumption observations for one and the same individual at low cost. This can yield a strong statistical test even when focusing on the standard utility maximization model. Finally, our application to real-life data illustrates the possibility of adding additional structure on the preferences of the consumer (in our case, quasi-linearity) to strengthen the test. If the observed behavior is (approximately) utility maximizing for such more structured preferences, it will generally be easier to statistically distinguish optimizing behavior from random behavior.

References

Afriat, S. N., 1967. The construction of utility functions from expenditure data. International Economic Review 8 (1), 67–77.

Andreoni, J., Miller, J., 2002. Giving according to garp: An experimental test of the consistency of preferences for altruism. Econometrica 70, 737–753.

Apesteguia, J., Ballester, M. A., 2015. A measure of rationality and welfare. Journal of Political Economy 123, 1278–1310.

Beatty, T. K. M., Crawford, I. A., 2011. How demanding is the revealed preference approach to demand. American Economic Review 101, 2782–2795.

Becker, G. S., 1962. Irrational behavior and economic theory. Journal of Political Economy 70, 1–13.

Bronars, S. G., 1987. The power of nonparametric tests of preference maximization. Econometrica, 693–698.

Brown, D. J., Calsimiglia, C., 2007. The nonparametric approach to applied welfare analysis. Economic theory 31, 183–188.

Chambers, C., Echenique, F., 2016. Revealed Preference Theory. Cambridge University Press.

Choi, S., Fisman, R., Gale, D., Kariv, S., 2007. Consistency and heterogeneity of individual behavior under uncertainty. American Economic Review 97, 1921–1938.

Choi, S., Kariv, S., Müller, W., Silverman, D., 2014. Who is (more) rational? The American Economic Review 104, 1518–1550.

Dean, M., Martin, D., 2016. Measuring rationality with the minimum cost of revealed preference violations. Review of Economics and Statistics 98, 524–534.

Diewert, W. E., 1973. Afriat and revealed preference theory. The Review of Economic Studies 40, 419–425.

Dziewulski, P., 2019. Just-noticeable difference as a behavioural foundation of the critical cost-efficiency index. Tech. rep., Working paper 05-2019, University of Sussex.

Echenique, F., Lee, S., Shum, M., 2011. The money pump as a measure of revealed preference violations. Journal of Political Economy 119, 1201–1223.

Fisman, R., Kariv, S., Markovits, D., 2007. Individual preferences for giving. American Economic Review 97, 1858–1876.

Fostel, A., Scarf, H. E., Todd, M. J., 2004. Two new proofs of Afriat's theorem. Economic Theory 24, 211–219.

Houthakker, H. S., 1950. Revealed preference and the utility function. Economica 17, 159–174.

Houtman, M., Maks, J. A. H., 1985. Determining all maximal data subsets consistent with revealed preference. Kwantitatieve Methoden 19, 89–104.

Kitamira, Y., Stoye, J., 2018. Nonparametric analysis of random utility models. Econometrica 86, 1883–1909.

Pesarin, F., Salmaso, L., 2010. The permutation testing approach: A review. Statistica 70, 481–509.

Polisson, M., Quah, J. K.-H., Renou, L., 2020. Revealed preferences over risk and uncertainty. American Economic Review, Forthcoming.

Samuelson, P. A., 1938. A note on the pure theory of consumer's behaviour. Economica 5, 61–71.

Selten, R., 1991. Properties of a measure of predictive success. Mathematical Social Sciences 21, 153–167.

Varian, H. R., 1982. The nonparametric approach to demand analysis. Econometrica, 945–973.

Varian, H. R., 1990. Goodness-of-fit in optimizing models. Journal of Econometrics 46, 125–140.

Varian, H. R., 1991. Goodness-of-fit for revealed preference tests. Tech. rep.

A Proofs

A.1 Proof of Theorem 1

Let $D = (q^t, p^t)_{t \leq T} \in E^{U,e}$. Then, for all $t \leq T$,

$$U(q^t) \geq \max_{\tilde{q}} U(\tilde{q}) \text{ s.t. } p^t \cdot \tilde{q} \leq e(p^t \cdot q^t).$$

Let $q^t R^e q^v$, i.e. $e(p^t \cdot q^t) \geq p^t \cdot q^v$. Then, $U(q^t) \geq \max_{\{\tilde{q}: p^t \cdot \tilde{q} \leq e(p^t \cdot q^t)\}} U(\tilde{q}) \geq U(q^v)$, so $U(q^t) \geq U(q^v)$. Similarly, we can show that $q^t P^e q^v$ implies $U(q^t) > U(q^v)$. Then, if e-GARP is violated, we have that there is a sequence $t_1, \ldots, t_M \leq T$ such that

$$q^{t_1} R^e \ldots R^e q^{t_M} \text{ and } q^{t_M} P^e q^{t_1}.$$

However, this implies

$$U(q^{t_1}) \geq \ldots \geq U(q^{t_M}) \text{ and } U(q^{t_M}) > U(q^{t_1}),$$

a contradiction.

For the reverse, let $D = (q^t, p^t)_{t \leq T}$ satisfy e-GARP. From Fostel, Scarf, and Todd (2004), we know that there exist numbers U^t and $\lambda^t > 0$ such that, for all observations $t, v \leq T$,

$$U^t - U^v \leq \lambda^v p^v \cdot (q^t - eq^v). \tag{2}$$

Consider the utility function

$$V(q) = \min_{t \leq T} \Big\{ U^t + \lambda^t p^t \cdot (q - eq^t) \Big\}.$$

The function V is increasing, continuous and concave. Let us first show that $V(q^t) \geq U^t$. If not, there must exist an observation v such that

$$V(q^t) = U^v + \lambda^v (q^t - eq^v) < U^t.$$

However, this contradicts equation (2). Now, towards a contradiction, assume that $D \notin E^{V,e}$. Then, there is an observation t such that

$$V(q^t) < \max_{\tilde{q}} V(\tilde{q}) \text{ s.t. } p^t \cdot \tilde{q} \leq e(p^t \cdot q^t).$$

Let $\tilde{q}^*$ solve the maximization problem on the right hand side. Then,

$$\begin{aligned} U^t &\leq V(q^t) \\ &< V(\tilde{q}^*) \\ &\leq U^t + \lambda^t p^t \cdot (\tilde{q}^* - eq^t) \\ &= U^t + \lambda^t (p^t \cdot \tilde{q}^* - e(p^t \cdot q^t)) \\ &\leq U^t, \end{aligned}$$

a contradiction.

A.2 Proof of Theorem 2

Fix $\alpha \in (0,1)$ and let $P \in \mathcal{P}_0$. Then, given the definition of $\mathcal{P}_0$, it is clear that under P the random data set D and the permuted data set D_σ have the same distribution: for all measurable sets $A \in \mathcal{B}$ and all permutations $\sigma \in \Pi$,

$$P(D \in A) = P(D_\sigma \in A).$$

As such, we have that, for all $\sigma \in \Pi$,

$$\mathbb{E}_P(\phi_\alpha(D)) = \mathbb{E}_P(\phi_\alpha(D_\sigma)),$$

and

$$\begin{aligned}\mathbb{E}_P(\phi_\alpha(D)) &= \frac{1}{T!}\sum_{\sigma\in\Pi}\mathbb{E}_P(\phi_\alpha(D_\sigma)) \\ &= \frac{1}{T!}\mathbb{E}_P\left[\sum_{\sigma\in\Pi}\phi_\alpha(D_\sigma)\right],\end{aligned}$$

where the last equality follows from exchanging integration and summation. For the sum within the expectation sign, we have

$$\sum_{\sigma\in\Pi}\phi_\alpha(D_\sigma) = \sum_{\sigma\in\Pi}\mathbf{1}\left[\frac{\left|\left\{\rho\in\Pi : \tau(D_\rho) \geq \tau(D_\sigma)\right\}\right|}{T!} \leq \alpha\right].$$

Now consider a ranking of all data sets D_σ for $\sigma \in \Pi$ according to their CCEI, $\tau(D_\sigma)$, from smallest to largest. Then, $\phi_\alpha(D_\sigma)$ will be zero for the lowest values of the ranking and will be equal to 1 from the $(1-\alpha)$th quantile onward. As such, we have

$$\sum_{\sigma\in\Pi}\phi_\alpha(D_\sigma) \leq \alpha\,(T!).$$

Combining all this, we obtain

$$\begin{aligned}\mathbb{E}_P(\phi_\alpha(D)) &\leq \frac{1}{T!}\mathbb{E}_P\left[(T!)\alpha\right] \\ &= \alpha.\end{aligned}$$

A.3 Proof of Theorem 3

Let $e \in [0,1]$ and $\widehat{P}_T \in \widehat{\mathcal{P}}^e_{T,1}$. Let $(\Pi, 2^\Pi, Q)$ denote the uniform probability space on Π, i.e. Q is the probability measure on $(\Pi, 2^\Pi)$ such that, for all $S \subseteq \Pi$,

$$Q(S) = \frac{|S|}{T!}.$$

Our aim is to construct an upper bound on $\mathbb{E}_{\widehat{P}_T}(1 - \phi_\alpha(D))$, i.e. the probability of not rejecting the null hypothesis. Notice that, for a random data set D of size T,

$$1 = \widehat{P}_T\left(D \in \left(\widehat{E}^{U,e}\right)^T\right) \leq \widehat{P}_T(D \text{ satisfies } e\text{-GARP}) \leq \widehat{P}_T(\tau(D) \geq e).$$

As such,

$$\begin{aligned}
\mathbb{E}_{\widehat{P}_T}(1 - \phi_\alpha(D)) &= \mathbb{E}_{\widehat{P}_T}(1 - \phi_\alpha(D) | \tau(D) \geq e) \\
&\leq \mathbb{E}_{\widehat{P}_T}\left[\frac{|\sigma \in \Pi : \tau(D_\sigma) \geq \tau(D)|}{|\Pi|} > \alpha \middle| \tau(D) \geq e\right] \\
&\leq \mathbb{E}_{\widehat{P}_T}\left[\frac{|\sigma \in \Pi : \tau(D_\sigma) \geq e|}{|\Pi|} > \alpha\right] \\
&\leq \mathbb{E}_{\widehat{P}_T}\left[\mathbb{E}_Q\left[\mathbf{1}\left(\tau(D_\sigma) \geq e\right)\right] \geq \alpha\right].
\end{aligned}$$

Next, applying Markov's inequality gives

$$\begin{aligned}
\mathbb{E}_{\widehat{P}_T}(1 - \phi_\alpha(D)) &\leq \frac{1}{\alpha}\mathbb{E}_{\widehat{P}_T}\big[\mathbb{E}_Q\left[\mathbf{1}\left(\tau(D_\sigma) \geq e\right)\right]\big] \\
&= \frac{1}{\alpha}\mathbb{E}_Q\big[\mathbb{E}_{\widehat{P}_T}\left[\mathbf{1}\left(\tau(D_\sigma) \geq e\right)\right]\big],
\end{aligned}$$

where the last equality follows from exchanging integration and summation.

We say that a permutation $\sigma \in \Pi$ has n fixed points if there are exactly n integers $i \in \{1, \ldots, T\}$ such that $\sigma(i) = i$. Let Π_n be the set of permutations with n fixed points. Then,

$$
\begin{aligned}
\mathbb{E}_{\widehat{P}_T}(1-\phi_\alpha(D)) &\leq \frac{1}{\alpha}\mathbb{E}_Q\left[\mathbb{E}_{\widehat{P}_T}\left[\mathbf{1}\left(\tau(D_\sigma)\geq e\right)\right]\right] \\
&= \frac{1}{\alpha}\sum_{n=0}^{T} Q(\Pi_n)\sum_{\sigma\in\Pi_n}\frac{1}{|\Pi_n|}\mathbb{E}_{\widehat{P}_T}\left[\mathbf{1}\left(\tau(D_\sigma)\geq e\right)\right] \\
&\leq \frac{1}{\alpha}\sum_{n=0}^{T} Q(\Pi_n)\sum_{\sigma\in\Pi_n}\frac{1}{|\Pi_n|}\pi_e^{\frac{T-n}{4}} && \text{(by Lemma 1)} \\
&\leq \frac{1}{\alpha}\sum_{n=0}^{T}\frac{1}{n!}\pi_e^{\frac{T-n}{4}} && \text{(by Lemma 2)} \\
&\leq \frac{1}{\alpha}\pi_e^{\frac{T}{4}}\sum_{n=0}^{\infty}\frac{\pi_e^{-\frac{n}{4}}}{n!} \\
&= \frac{1}{\alpha}\pi_e^{\frac{T}{4}}\exp\left(\pi_e^{-\frac{1}{4}}\right) \\
&\leq \frac{1}{\alpha}\pi_e^{\frac{T}{4}}\exp\left(\left(\frac{1}{2}\right)^{-\frac{1}{4}}\right) \leq \frac{4}{\alpha}\pi_e^{\frac{T}{4}},
\end{aligned}
$$

which completes the proof.

A.4 Lemmata

Lemma 1. *Let $\sigma \in \Pi$ and $\widehat{P}_T \in \widehat{\mathcal{P}}^e_{T,1}$. Further, let D be a data set that is drawn randomly from $(\widehat{\Omega}_T, \widehat{\mathcal{B}}_T, \widehat{P}_T)$. Let $n = |\{i \leq T : \sigma(i) = i\}|$ be the number of fixed points of σ. Then,*

$$\widehat{P}_T(\tau(D_\sigma)\geq e) \leq \pi^{\frac{T-n}{4}},$$

where $\pi \leq \max(\widetilde{\pi}_e, \overline{\pi}_e)$.

Proof. Any permutation can be decomposed into an exhaustive set of disjoint cycles. As there are T elements in total in the set $\{1,\ldots,T\}$, the maximum length of a cycle in σ is T. Also, the statement of the lemma implies n cycles of length 1. Let us denote by C_m the number of cycles in the permutation σ of length m. Then, calculating the elements by cycle gives

$$T = \sum_{k=1}^{T} kC_k = n + \sum_{k=2}^{T} kC_k. \tag{3}$$

Consider a cycle of length 2. As this cycle has no fixed points, it must take the form

$$\widetilde{D} = \Big((r^{\sigma(i)}, p^i, m^i), (r^i, p^{\sigma(i)}, m^{\sigma(i)})\Big),$$

for some $i \leq T$. As observations are i.i.d., each such data subset $\widetilde{D}$ generated from a cycle of length 2 must be independent of all other observations in D.

Next, any cycle of length $k \geq 3$ allows for constructing $\lfloor \frac{k}{3} \rfloor$ non-overlapping data sets of size 3 that take the form

$$\left(\left(r^{\sigma(i)}, p^i, m^i\right), \left(r^{\sigma(\sigma(i))}, p^{\sigma(i)}, m^{\sigma(i)}\right), \left(r^{\sigma(\sigma(\sigma(i)))}, p^{\sigma(\sigma(i))}, m^{\sigma(\sigma(i))}\right)\right),$$

for some $i \leq T$ and $\lfloor a \rfloor$ denoting the greatest integer below a. Consider the data subsets generated from these three element data sets by dropping the last observation,

$$\overline{D} = \left(\left(r^{\sigma(i)}, p^i, m^i\right), \left(r^{\sigma(\sigma(i))}, p^{\sigma(i)}, m^{\sigma(i)}\right)\right).$$

All these data sets $\overline{D}$ are independent of all other data subsets that are constructed from the same cycle (as they have no indices in common), and they are independent of observations belonging to other cycles. The total number of both types of independent data subsets of size 2 is then bounded from below by

$$\begin{aligned} C_2 + \sum_{k=3}^{T} C_k \left\lfloor \frac{k}{3} \right\rfloor &\geq C_2 + \sum_{k=3}^{T} C_k \frac{k}{4} \\ &\geq \frac{1}{4} \sum_{k=2}^{T} k\, C_k \\ &= \frac{1}{4}(T - n). \qquad \text{(by equation (3))} \end{aligned}$$

The first inequality follows from the fact that, for $k \geq 3$, $\lfloor k/3 \rfloor \geq k/4$. If $\tau(D_\sigma) \geq e$, then all these independent data sets also have CCEI values ($\tau(\widetilde{D})$ and $\tau(\overline{D})$) that exceed e. As such, the probability that $\tau(D_\sigma) \geq e$ is bounded from above by

$$\widehat{P}_T(\tau(D_\sigma) \geq e) \leq \prod_{i=1}^{\frac{T-n}{4}} \pi = \pi^{\frac{T-n}{4}}.$$

□

Lemma 2. *We have*

$$Q(\Pi_n) \leq \frac{1}{n!}. \tag{4}$$

Proof. A derangement is defined as a permutation with no fixed points. Let $!n$ be the number of derangements of $\{1,\dots,n\}$. It directly follows that $!n/n! \leq 1$ and, thus,

$$Q(\Pi_n) = \frac{\binom{N}{n}\big(!(N-n)\big)}{N!} = \frac{!(N-n)}{n!((N-n)!)} \leq \frac{1}{n!}.$$

Here, we counted the number of elements in Π_n by first counting the possible ways to pick n fixed points and then, for each set of fixed points, counting the possible ways in which the remaining elements can be deranged. □

B Algorithm to generate random data sets that satisfy e-GARP

For an observation (q^t, p^t) we define the share vector

$$s_i^t = \frac{p_i^t q_i^t}{m^t},$$

where m^t represents the expenditure level at the observation, i.e. $m^t = \sum_{i=1}^L p_i^t q_i^t$. As there is a one-to-one correspondence between (q^t, p^t) and (s^t, p^t, m^t), we will interchangeably denote a data set $D = (q^t, p^t)_{t\leq T}$ as $D = (s^t, p^t, m^t)_{t\leq T}$.

Our algorithm is a Markov-Chain-Monte-Carlo (MCMC) procedure that is based on combining a Gibbs sampler with a hit-and-run step. Starting from a data set D that satisfies e-GARP, at each iteration we pick one observation $t \in \{1,\dots,T\}$ at random, and we update the share vector s^t for observation t using a hit-and-run over the set of all share vectors s^t that preserve e-GARP consistency.

More specifically, consider a data set $D = (s^t, p^t, m^t)_{t\leq T}$. We use the following notation for the new data set that is obtained by replacing s_t in this original data set by the vector $\tilde{s}^t$:

$$(\tilde{s}^t, s^{-t}) = (s^1, \dots, s^{t-1}, \tilde{s}^t, s^{t+1}, \dots, s^T).$$

If D is consistent with e-GARP, we define

$$\mathcal{P}(s^{-t}) = \{\tilde{s}^t \in \Delta^L : (\tilde{s}^t, s^{-t}) \text{ satisfies } e\text{-GARP}\}.$$

The set $\mathcal{P}(s^{-t})$ contains all share vectors $\tilde{s}^t$ in the L-dimensional simplex Δ^L such that the data set obtained by replacing s^t by $\tilde{s}^t$ still satisfies e-GARP. It can be shown that $\mathcal{P}(s^{-t})$ is a convex set.

Let us first describe our hit-and-run routine. We use ∂S to denote the set of all

Algorithm 1 Sample δ uniformly from ∂S

Require: L

1: Draw L i.i.d. standard normally distributed variables $x_1, \ldots, x_L$
2: Compute $y = (y_1, \ldots y_L)$ where $y_i \leftarrow x_i - \sum_{i=1}^{L} \frac{x_i}{L}$
3: Compute $\delta = (\delta_1, \ldots, \delta_L)$ where $\delta_i \leftarrow \frac{y_i}{\|y\|}$
4: **return** δ

directions inside the L-dimensional simplex Δ^L, i.e.

$$\partial S = \left\{ \delta \in \mathbb{R}^L : \|\delta\| = 1 \text{ and } \sum_i \delta_i = 0 \right\}.$$

The set ∂S contains all vectors whose elements sum to zero that are on the surface of the L-dimensional unit sphere. Algorithm 1 shows how to draw an element δ uniformly from ∂S.

For a given share vector $s \in \Delta^L$ and direction $\delta \in \partial S$, let

$$\overline{\lambda} = \sup_{\lambda} \{ s + \lambda\delta \in \Delta^L \}.$$

This value $\overline{\lambda}$ can easily be determined. Observe that if s^t is a share vector and $\delta \in \partial S$, then we have $\sum_{i=1}^{L}(s_i + \lambda\delta_i) = 1$ for any $\lambda \in \mathbb{R}$. As such, for all goods $i = 1, \ldots, L$, it suffices to just consider those λ for which

$$0 \leq s_i + \lambda\delta_i \leq 1.$$

If $\delta_i > 0$, this gives the bounds

$$\lambda \leq \frac{1 - s_i}{\delta_i},$$

and if $\delta_i < 0$, we obtain the bounds

$$\lambda \leq \frac{-s_i}{\delta}.$$

Since not all δ_i can be zero, we thus have

$$\overline{\lambda} = \min \left\{ \min_i \left\{ \frac{1 - s_i}{\delta_i} : \delta_i > 0 \right\}, \min_i \left\{ \frac{-s_i}{\delta_i} : \delta_i > 0 \right\} \right\}.$$

Next, for a given data set $D = (s^t, p^t, m^t)_{t \leq T}$ and direction δ, we define

$$\tilde{\lambda}^t = \sup\{\lambda \leq \overline{\lambda^t} : (s^t + \lambda\delta, s^{-t}) \in \mathcal{P}(s^{-t}).\}$$

The value of $\tilde{\lambda}^t$ can be found through a binary search routine, as shown in Algorithm 2.

Algorithm 2 Compute $\tilde{\lambda}^t$ up to an error ε

Require: A data set $D = (s^t, p^t, m^t)_{t \leq T}$ that satisfies e-GARP, an observation $t \leq T$, a direction $\delta \in \partial S$ and an error term $\varepsilon > 0$

compute $\overline{\lambda^t}$
$a \leftarrow \overline{\lambda^t}$
$b \leftarrow 0$
while $(a - b) \geq \varepsilon$ **do**
  $c \leftarrow \frac{(b+a)}{2}$
  if $(s^t + c\delta, s^{-t})$ satisfies e-GARP **then**
    $b \leftarrow c$
  else
    $a \leftarrow c$
  end if
end while
return $\tilde{\lambda}^t = c$

Finally, Algorithm 3 provides the updating step of our MCMC procedure. To get a new data set that satisfies e-GARP from a given set $(s^t)_{t \leq T}$ that is consistent with e-GARP (when fixing the prices and expenditure levels), we first draw a permutation σ on $\{1, \ldots, T\}$ at random, and we use the hit-and-run step to sequentially update the share vectors $s^{\sigma(1)}, \ldots, s^{\sigma(T)}$ to their new values. This is done by, first, drawing a random direction δ uniformly from ∂S, next, computing the value of $\tilde{\lambda}^{\sigma(t)}$ and, finally, drawing a value λ uniformly from $(0, \tilde{\lambda}^{\sigma(t)})$. The new value of $s^{\sigma(t)}$ is set equal to $s^{\sigma(t)} + \lambda\delta$. The requirement that $\lambda \in (0, \tilde{\lambda}^{\sigma(t)})$ together with convexity of the set $\mathcal{P}(s^{-t})$ guarantees that every updated data set also satisfies e-GARP.

To intialize the MCMC procedure, we need a vector of shares $(s^t)_{t \leq T}$ that is consistent with e-GARP (for given prices and expenditure levels). To this end, we choose expenditure shares such that a fraction $1/L$ of the total budget is spent on each good. This corresponds to optimizing behavior for a Cobb-Douglas utility function that attaches equal weights to all goods, which guarantees consistency with e-GARP for any $e \in [0, 1]$.

Algorithm 3 Generate M random data sets that satisfy e-GARP given an initial data $D = (s^t, p^t, m^t)_{t \leq T}$ that satisfies e-GARP

Require: A data set $D = (s^t, p^t, m^t)_{t \leq T}$ that satisfies e-GARP

Initialize $n \leftarrow 0$
while $n \leq M$ **do**
  Randomly pick a permutation $\sigma : \{1, \ldots, T\} \to \{1, \ldots, T\}$
  for $t = 1$ **to** T **do**
    Draw a direction δ uniformly from ∂S using Algorithm 1
    Compute $\tilde{\lambda}^{\sigma(t)}$ using Algorithm 2 for the observation $\sigma(t)$ and the direction δ
    Draw λ uniformly from $(0, \tilde{\lambda}^{\sigma(t)})$
    $s^{\sigma(t)} \leftarrow s^{\sigma(t)} + \lambda \, \delta$
  end for
  $n \leftarrow n + 1$
end while