**Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
Bonn 2019/1**

Estimating Heterogeneous Reactions to Experimental Treatments

Christoph Engel

MAX PLANCK SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods

Bonn 2019/1

Estimating Heterogeneous Reactions to Experimental Treatments

Christoph Engel

January 2019

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Estimating Heterogeneous Reactions to Experimental Treatments[*]

Christoph Engel

This version: January 20, 2019

Abstract

Frequently in experiments there is not only variance in the reaction of participants to treatment. The heterogeneity is patterned: discernible types of participants react differently. In principle, a finite mixture model is well suited to simultaneously estimate the probability that a given participant belongs to a certain type, and the reaction of this type to treatment. Yet often, finite mixture models need more data than the experiment provides. The approach requires ex ante knowledge about the number of types. Finite mixture models are hard to estimate for panel data, which is what experiments often generate. For repeated experiments, this paper offers a simple two-step alternative that is much less data hungry, that allows to find the number of types in the data, and that allows for the estimation of panel data models. It combines machine learning methods with classic frequentist statistics.

Keywords: heterogeneous treatment effect, finite mixture model, panel data, two-step approach, machine learning, CART

JEL Classification: C14, C23, C91

[*] Prof. Dr. Christoph Engel, Max Planck Institute for Research on Collective Good, D 53173 Bonn, engel@coll.mpg.de, ++49 228 91416210, ORCID 0000-0001-8513-8532. Helpful comments by Sebastian Schneider, Marcel Schubert and audiences at the Economics Department of National University of Taiwan, and at the Max Planck Institute for Research on Collective Goods at Bonn, are gratefully acknowledged.

1 Introduction

Not all experimental participants are equal. This is not only a truism. If, strictly speaking, all were equal, there would be nothing to estimate. There would be no need to expose a randomly selected sample to random variation. One could infer the universal law of nature from exposing one of two otherwise identical individuals to treatment. Most empirical researchers shy away from the philosophical debate over natural laws. Even if such laws exist, and matter for the behavior of human participants, the researcher is not in a position to observe them. All she can study is the reaction of a sample that she suspects to differ in multiple ways. Yet as long as (a) assignment to treatment is random, (b) the sample is randomly drawn from the population and sufficiently large, and (c) reactions to treatment are sufficiently pronounced, the researcher can infer the population effect. Frequentist statistics let her assess whether it is sufficiently unlikely for the observed difference to be a false positive.

Note that this standard approach to the analysis of experimental data assumes heterogeneity: different individuals react differently to treatment. Yet this heterogeneity is treated as a nuisance variable. It results from the fact that perfectly clean data is unavailable. It is the purpose of randomization to prevent this heterogeneity from biasing the estimation of the treatment effect. The researcher feels justified to treat the unobserved heterogeneity as noise. This is why, in statistical textbooks, the estimation of the treatment effect is introduced as the difference in the central tendency of two Gaussian distributions.

Not so rarely, experimenters have reason to doubt that the heterogeneity in reaction to treatment is indeed random. A prominent illustration is social preferences. On average, participants in dictator, ultimatum or public good games do not behave as predicted by microeconomic textbooks. They share some of their endowments with their passive counterparts (Engel, 2011), they reject offers that exploit a first-mover advantage (cf. Cooper and Dutcher, 2011), and they make substantial contributions to socially beneficial joint projects (Zelmer, 2003). Yet a substantial fraction of most experimental samples maximize short-term profit. A rather small minority are true altruists. And many only neglect the dilemma structure of a public good if they know, observe or believe that their counterparts will do so as well (Fischbacher et al., 2001).[1] There are thus (at least) three discernible types.

In principle, such patterned heterogeneity is a case for a finite mixture model. The model can be estimated with maximum likelihood. The procedure simultaneously estimates the probability of a datapoint to belong to each of the types, and the reaction of each type to treatment. In postestimation, each participant of the experiment can be assigned to the most likely type. Yet in practice, finite mixture models with experimental data often do not converge. Two-dimensional maximum likelihood requires more and cleaner data than many experiments produce. This in particular holds if one prop-

[1] For detail see below Section 5.

erly reflects the dependence structure induced by an interactive, repeated experiment, like a repeated public good. Such an experiment generates data from choices, nested in individuals, nested in groups. Each group is a single independent observation. A further drawback is the necessity to fix the number of types beforehand, although one typically estimates the finite mixture model when the data suggests that there might be more than a single type. The purpose of the model is to find out whether heterogeneity is indeed patterned. At this point of the research process, theory to rationalize the type space is still missing.

In this paper, I propose a statistical approach that precisely targets this situation. With the proposed approach, the biggest challenge for the estimation of a finite mixture model turns into the critical asset: the panel structure of repeated experiments. The approach needs one identifying assumption: type is a personality variable. The heterogeneity originates in the fact that different individuals react to treatment in different ways. If this assumption can be made, in a first step one can separately regress each individual on all (time-varying) independent variables. The coefficients from these local regressions characterize the individual's type. Standard machine learning techniques, like a classification and regression tree CART,[2] can be used to find the best way to partition the type space. In the second step, each participant is characterized by one of these types. If treatment is within subjects, this procedure directly produces estimates of the treatment effect conditional on type. If treatment is (exclusively) between subjects, the procedure separates the reactions of different types of subjects to treatment. Yet one needs additional assumptions (or complementary within subjects data) to match types of untreated with types of treated subjects. If one has reason to trust the matching, one can interact treatment with type. The interaction terms then estimate in which ways the reactions of different types to treatment differ.

The remainder of the paper is organized as follows. The next section relates the paper to the literature. Section 3 explains the approach in detail, and contrasts it with the alternatives. Section 4 uses simulation to explore how well the proposed non-parametric method performs. Section 5 applies the approach to a real experimental dataset. Section 6 concludes with discussion.

2 Related Literature

Experimenters pay increasingly attention to patterned heterogeneity (see, for instance, Bruhin et al., 2018; Conte and Levati, 2014; Santos-Pinto et al., 2015) and use finite mixture models (Moffatt, 2015) to simultaneously estimate the composition of the type space, and reactions to treatment conditional on type, in games as diverse as public goods (Bardsley and Moffatt, 2007; Kassas et al., 2018), prisoner dilemmas (Becchetti

[2] The logic of CART is explained below. The code for implementing CART is available in the technical appendix.

et al., 2017), beauty contests (Bosch-Domènech et al., 2010; Breitmoser, 2012), bribery games (Bolle et al., 2011), learning in networks (Kovářík et al., 2018), and attitudes towards macro-risk in financial markets (Brown and Kim, 2013). Yet to the best of my knowledge, none of these papers discuss machine learning methods to organize the type space.

There is an active literature on the estimation of heterogeneous treatment effects outside experimental economics. Some of these papers discuss the application of machine learning methods (for overviews see Alaa and Schaar, 2018; Künzel et al., 2017; Powers et al., 2017). They for instance use CART (Athey and Imbens, 2016; Su et al., 2009), random forests (Lu et al., 2018; Wager and Athey, 2017) or support vector machines (Imai et al., 2013) to estimate differences in the reaction to treatment, or advocate averaging types over the outcomes from multiple alternative machine learning methods (Grimmer et al., 2017).

A particularly active application is biostatistics. Data from reactions of patients to alternative medical interventions is used to personalize treatment (Bonetti and Gelber, 2004; Gail and Simon, 1985; Sauerbrei et al., 2007; Tian et al., 2014; Wendling et al., 2018; Zhao et al., 2012) or to evaluate the performance of hospitals in treating a heterogeneous population of patients (Berta et al., 2016).

Closest in spirit is Bonhomme et al. (2016). They also propose to proceed in two steps. In the first step, they estimate the probability that a datapoint belongs to a certain group, exploiting repeated measurement. In the second step, the data are weighted by these estimates. Yet they assume the number of groups (types) to be known ex ante, while my approach allows to estimate them from the data. Bertoletti et al. (2015) propose a Bayesian method to estimate the number of groups in a finite mixture from the data. As I will explain below, under suitable conditions, there is a simpler approach if one has multiple observations per participant of an experiment.

3 Estimation Approaches

Observed Type

If the type space is fully understood, a two-step approach invites itself. In a first step, one measures type, for instance with the test developed by Fischbacher et al. (2001). In a second step, one explains observed choices y_i with type $\tau_i \in \{1, .., T\}$ and treatment $\theta_i \in \{0, 1\}$. Hence one estimates

$$y_i = \begin{cases} \beta_{1,0} + \beta_{1,1}\theta_i + \epsilon_i \quad \text{if} \quad \tau_i = 1, \\ ... \\ \beta_{T,0} + \beta_{T,1}\theta_i + \epsilon_i \quad \text{if} \quad \tau_i = T. \end{cases} \tag{1}$$

This can equivalently be written as

$$y_i = \beta_0 + \beta_1\theta_i + \sum_{\tau=2}^{T} \beta_\tau \tau_i + \sum_{\tau=2}^{T} \beta_{T-1+\tau}\tau_i \cdot \theta_i + \epsilon_i. \tag{2}$$

One defines one type as the reference category. For this type, β_0 is the estimated choice when untreated, and $\beta_0 + \beta_1$ is the estimated choice when treated. For any other type, the choice when untreated is estimated by $\beta_0 + \beta_\tau$, and the choice when treated is estimated by $\beta_0 + \beta_1 + \beta_\tau + \beta_{T-1+\tau}$. This specification has the advantage that $\beta_2..\beta_T$ are a direct estimate for the difference between the type chosen as the reference category and the respective alternative type when untreated. Likewise the interaction terms measure how the reaction to treatment differs between the reference type and the remaining types.[3]

Finite Mixture Model

If type τ_i is not observed independently of choice y_i, the composition of the type space, and choices conditional on type, must be simultaneously estimated. In principle, this can be done with a finite mixture model. If one feels confident to estimate a linear model, the density to be estimated is given by

$$f(y_i) = \sum_{\tau=1}^{T} \pi_\tau f_\tau(y_i|\boldsymbol{x_i}'\boldsymbol{\beta}). \tag{3}$$

In (3) $f_\tau(y_i|\boldsymbol{x_i}'\boldsymbol{\beta})$ is a generic way of writing (1), while allowing $\boldsymbol{x_i}$ to contain further covariates, together with the treatment variable θ_i. Yet through π_τ, the model allows for different types to react differently to treatment, and estimates the probability of an observation to be of a certain type, given independent variables $\boldsymbol{x_i}$ and the dependent variable y_i, with the constraint that $\sum_{\tau=1}^{T} \pi_\tau = 1$.

The model defined in (3) can be estimated with maximum likelihood. The probabilities $\pi_1..\pi_T$ are treated as latent variables. Estimating these latent variables is a challenge though. Statistical packages usually parry the challenge iteratively, using the EM algorithm (Dempster et al., 1977), going back and forth between (initially arbitrary) probabilities, and the coefficients, conditional on an observed datapoint belonging to one of the types.

This iterative procedure is why finite mixture models often fail with experimental data. Models do not converge. This is the more likely the more types one posits to exist. Moreover, in a finite mixture model, the number of types in the population must be fixed ex ante; it is not taken from the data.

[3] Of course both only holds if the statistical model is linear.

Estimating the Type Space from the Data

The previous approaches have treated each data point as an independent observation. Economic experiments are frequently repeated. For estimating the treatment effect, this is not a concern. One can estimate the random effects model (4):

$$y_{it} = \beta_0 + \beta_1 \theta_{it} + \epsilon_i + \epsilon_{it}, \tag{4}$$

assuming that individuals i are randomly exposed to treatment $\theta \in \{0, 1\}$, and that choices are nested in individuals i. Yet finite mixture models for panel data are difficult. The individual specific error ϵ_i is itself a random latent variable. One would be forced to integrate out latent variables in two dimensions (types, and individuals). One way out is adding dummies for individuals to $\boldsymbol{x}$ in (3) (Deb and Trivedi, 2013).[4]

For the approach proposed here, the panel structure of the data is, to the contrary, not a challenge, but the critical asset. For the approach to work, one must feel confident to assume that type is a personality variable. The population subdivides into an (initially unknown) number of types. Each individual is permanently of one and the same type. It depends on type how the individual reacts to treatment. The approach finally requires that type induces some within participant variation. The archetypal illustration is a time trend that differs across types.

If these conditions are fulfilled, one can proceed in two steps. In the first step one defines the type space and assigns each individual in the sample to one of these types. In the second step, one estimates the treatment effect conditional on (estimated) type.

Steps 1-9 of the Algorithm proposed below explain in which ways the panel structure of the data can be exploited to estimate the type space from the data. This part of the procedure has two components. One first regresses the choices y_{it} of each individual on all time varying observed explanatory variables $\boldsymbol{x}_{it}$ (step 3 of the algorithm). This yields for every participant a series of coefficients $\boldsymbol{\beta}_i$. These coefficients characterize the between subjects variance in the data.

The second component uses these coefficients to organize the type space (steps 5-7 of the Algorithm). The purpose of the exercise is estimating a heterogeneous treatment effect. Consequently, supervised learning is appropriate. One trains a classification algorithm on choices y_{it}, as explained by the individual coefficients $\boldsymbol{\beta}_i$. In principle, one could use any classification algorithm for the purpose, including naive Bayes, nearest neighbor methods, support vector machines or neural networks (for a very accessible introduction to these methods see James et al., 2013). Yet a classification tree CART is appealing for two reasons: the classification is straightforward to interpret, and there

[4] The workaround only works though if the panel is sufficiently long. Otherwise one runs into the incidental parameters problem (Neyman and Scott, 1948; Lancaster, 2000). And one inevitably looses information about the type specific reaction to observed characteristics that do not change over repetitions, as they are absorbed by the fixed effects.

are well-validated methods for defining the depth of the tree, and thereby the estimated number of types in the population (Breiman et al., 1984; Strobl et al., 2009).

CART recursively partitions the data, such that each split explains as much variance as possible. Hence at the first split, CART uses each coefficient in $\boldsymbol{\beta}$. As all coefficients are continuous, CART not only tries out each coefficient, but each cutpoint on each coefficient. This first step creates a tree with two branches. CART repeats the procedure and, separately for each branch of the tree, finds the (cutpoint at the) coefficient that explains most of the remaining variance. The standard CART algorithm first grows the complete tree, but then "prunes" it, to find the optimal balance between exploiting the information in the sample, and overfitting. The method proposed here uses this approach to find the optimal number of types. The problem is equivalent, as one only has the sample to estimate the type space in the population. Hence one has reason to be concerned about putting too much stress on unsystematic features of the sample. One needs to strike a balance between underusing and overusing the information present in the sample.

A tree that yields three types might for instance have a first split at $\beta_1 < 2$, and a second split for the right branch of the tree at $\beta_2 < 5$. These splits can also be used to assign participants to types. All participants with $\beta_1 < 2$ are classified as τ_1. Participants with $\beta_1 \geq 2$ and $\beta_2 < 5$ are classified as τ_2, and participants with $\beta_1 \geq 2$ and $\beta_2 \geq 5$ are classified as τ_3.

One uses these estimated types to estimate the dependent variable conditional on type (step 10 of the Algorithm). If treatment is within subjects, this step also yields an estimate of treatment conditional on type. If treatment is (exclusively) between subjects, one needs supplementary information, or must make assumptions, for matching untreated and treated types (step 11 of the Algorithm). In the final step (step 12 of the Algorithm) treatment effects conditional on type can then be recovered by way of postestimation. One uses Wald tests to estimate the treatment effect, separately for each type.

As, in step 10 of the Algorithm, one can treat participants as if one had always known their type, it is easy to capture the dependence structure by splitting up the error into $\epsilon_i + \epsilon_{it}$, i.e. by estimating a random effects model. This is particularly helpful if, as often, the data not only comes from a repeated, but from a repeated interactive experiment. Then choices are nested in individuals who are themselves nested in groups g.[5] This dependence structure can be captured by $\epsilon_g + \epsilon_{gi} + \epsilon_{git}$, i.e. by a mixed statistical model that distinguishes between the "fixed" effects $\boldsymbol{x}$ and the series of (assumedly orthogonal) random error terms (where g stands for the group).

Algorithm

[5] If groups are rematched during the experiment, g must stand for the matching group from which the rematching is done.

```
1. Let $D_0$ be a panel with dependent variable $y_{it}$, and explanatory variables $\boldsymbol{x}_{it}$ that include treatment $\theta_i$ (which may differ over repetitions, i.e. may be $\theta_{it}$)
2. initialize $\boldsymbol{\beta}$
   For every participant do
3.     regress $y_{it}$ on all time varying $\boldsymbol{x}_{it}$
4.     collect participant $id$ and all $\boldsymbol{\beta}_i$ in separate data frame $D_1$
   EndFor
5. merge $D_1$ with $D_0$ on $id$
6. fit classification tree of $y_{it}$ on $\boldsymbol{\beta}$
7. use standard algorithm to define optimal depth of tree
8. use optimal tree to assign type to each participant
   If treatment is between subjects
9.     split estimated types into treated and untreated cases
   EndIf
10. estimate panel version of (2)
   If treatment is exclusively between subjects $\theta_i$
11.     match untreated and treated types
12.     use postestimation for estimating treatment effects conditional on type
   EndIf
```

4 Simulation

In this section, I show with simulated data how the approach performs. I am making the R script for the simulation and analysis publicly available, so that researchers can use the code to adapt the approach to their own experimental data. The simulation is for a between subjects treatment, to also demonstrate the additional steps needed in this case.

In the simulation, $N = 400$ individuals are observed for $T = 10$ periods each. Half of the individuals are treated ($\theta_i \in \{1, 2\}$),[6] and individuals are of types $\tau_i \in \{1..4\}$. Types

[6] In (5), θ_i is not coded as a dummy variable as otherwise all untreated observations would be identical.

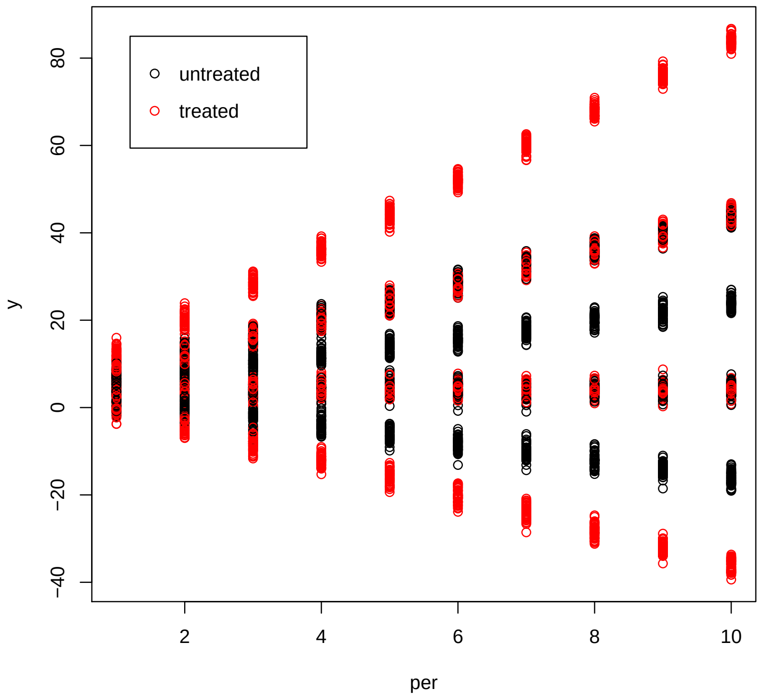

Figure 1: Simulated Data Pooled

differ in their reaction to treatment. Specifically, dependent variable y_{it} is generated according to (5)

$$y_{it} = 4 + 2 \cdot (3 - \tau_i) \cdot \theta_i \cdot t + \epsilon_i + \epsilon_{it}, \tag{5}$$

where individual specific error $\epsilon_i \sim \mathcal{N}(0, 1)$ captures dependence within individuals, and $\epsilon_{it} \sim \mathcal{N}(0, 1) \perp \epsilon_i$ is residual error. Figure 1 shows that the dependent variable seemingly exhibits 6 different groups. Both extremes come from treated data. Two intermediate arrows purely come from untreated participants. The remaining two arrows are mixed from treated and untreated participants (black and red circles overlap).

Comparing the regression in Table 1 with Figure 2 shows that ignoring the heterogeneity yields a very misleading picture. The regression finds overall a significant positive time trend. Yet this only holds for 2 of 4 types, while the trend is negative for type 4 and close to 0 for type 3. Likewise the interaction between treatment and the time trend is misleading. Overall it is again significantly positive. But this effect is driven by types 1 and 2, while the treatment effect is actually negative for type 4, and again close to 0 for type 3.

If this were experimental data, one would only have Figure 1. It clearly suggests

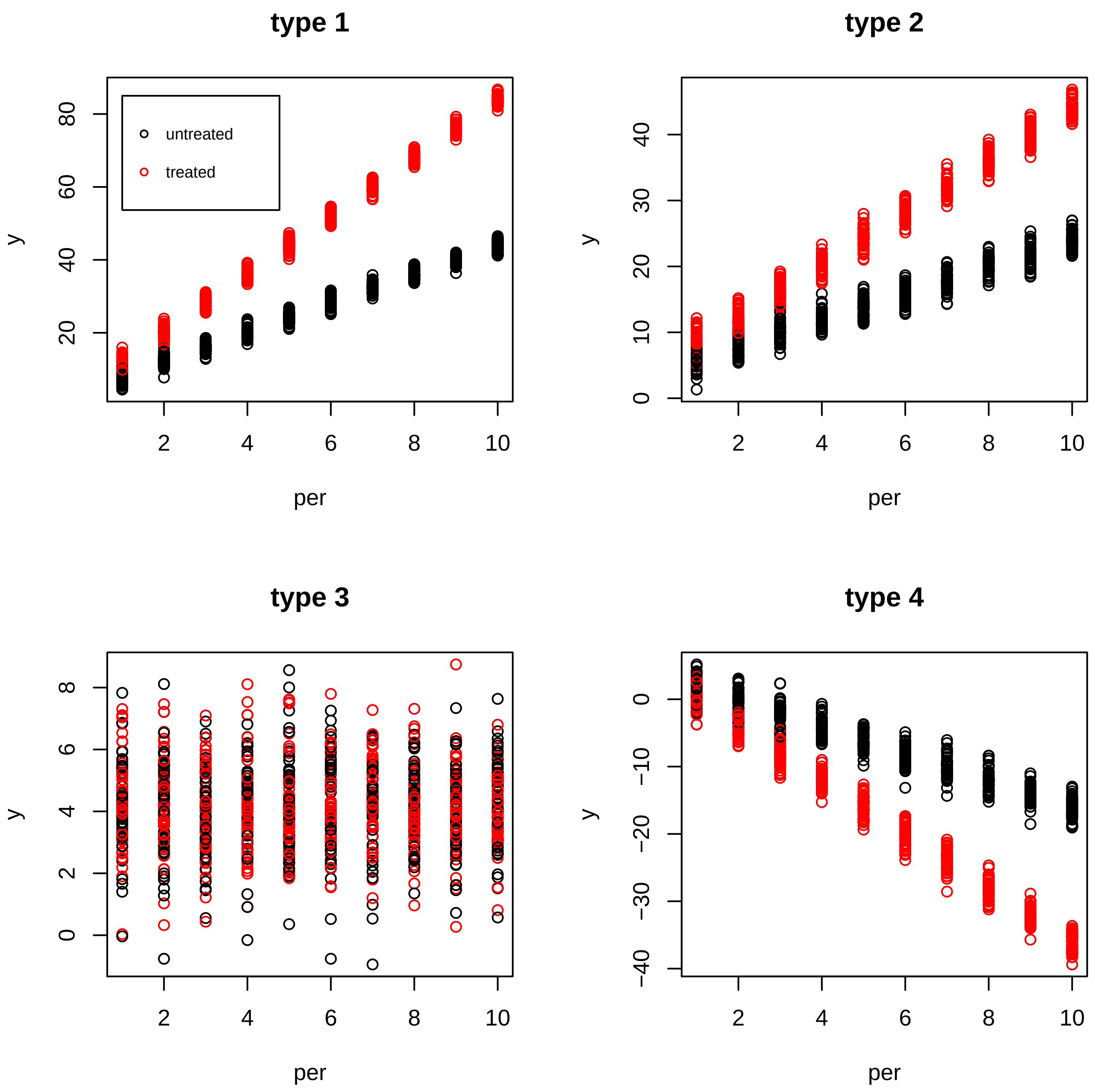

Figure 2: Simulated Data by Type

θ	0.105 (2.061)
t	0.994^{***} (0.084)
$\theta * t$	0.996^{***} (0.118)
cons	4.004^{**} (1.458)
N	4,000

Linear model with individual random effect. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

Table 1: Pooled Random Effects Model

patterned heterogeneity. But it is hard to guess the number of types: two, as there are some with a positive and some with a negative trend? Three, as there are two arrows that clearly separate untreated and treated cases? Or four, as is indeed the data generating process?[7]

As the simulated data generation process is so clean, the correct finite mixture model with four types converges and yields results that closely match (5), Table 2. Yet note that the model ignores dependence at the individual level.[8] As the data is simulated, I can compare estimated with true type. The estimate is correct in 96.33 % of all cases. The root mean squared error is 1.410, which is even less than in the original data, where it is 1.429.[9]

	type 1	type 2	type 3	type 4
p_τ	25.35	24.79	24.89	24.97
t	.055	.024	.012	-.007
θ	.301	.407	.147	-.102
$\theta * t$	3.963	1.982	-.011	-1.992
cons	3.548	3.367	3.949	4.002

Linear finite mixture model, assuming 4 groups, and treating all data points as independent

Table 2: Finite Mixture Model

[7] In an experiment, random assignment would exclude 6 types, as there could not be selection of types into treatment.

[8] Given each participant is assigned to either baseline or treatment for the entire sequence, one can also not emulate a fixed effects model by adding participant dummies: they would be perfectly collinear with the explanatory variable of interest, i.e. the interaction between τ and θ.

[9] This betrays a slight degree of overfitting: the finite mixture model "explains" some of the noise in the sample.

	1	2	3	4	5	6	7	8	9
untreated	50		60			40		50	
treated		50		8	50		42		50

Table 3: Type Space Estimated with CART

I now contrast this result with the result generated applying the Algorithm. The only variable that varies within participants is time $t = 1..10$. I therefore, separately for each individual, estimate

$$y_t = \beta_0 + \beta_1 t + \epsilon_t. \tag{6}$$

This step yields a new dataset with two scores per participant, β_0 and β_1, plus the observed outcomes y_{it}, and a user identifier. I use these scores to build the regression tree of Figure 3.[10] Two things are remarkable: the tree exclusively uses β_1, i.e. the individual slope coefficients, and it finds 6 types, i.e. the six distinct arrows of Figure 1.

Now θ is observed as well. If a type proposed by CART encompasses treated and untreated cases, one must split it up. Actually three of the types generated by CART exclusively cover treated or untreated cases. Taking this into account, the final set of types consists of nine types, of which four are treated, and 5 are untreated. Table 3 reports the estimated frequencies of these types.

Using this coding, in the next step I estimate (7), where $\hat{\tau}_k$ is one of the 9 estimated types.

$$y_{it} = \gamma_0 + \gamma_1 t_{it} + \sum_{k=2}^{9} \gamma_k \hat{\tau}_k + \sum_{k=2}^{9} \gamma_{2k} \hat{\tau}_k \cdot t_{it} + \epsilon_i + \epsilon_{it}. \tag{7}$$

Table 4 shows that the procedure works well. Type main effects are all insignificant, as they should, given the data generating process of (5) starts at the same point, irrespective of type. The coefficient of t captures the time trend for the first type (it corresponds to type 1 in Figure 2, for the untreated participants). The interaction effects define how much the time trend for each of the remaining estimated types $\hat{\tau}_k$ differs from the time trend in the first type.

In the simulation, treatment is exclusively between subjects. The algorithm exclusively uses slopes (β_1) for classification (as it should, given type does not affect the intercept). I assume that types are characterized by the proximity of slopes. This implies that type is assumed to be more important than treatment. Personality is the dominant factor, which is only moderated by treatment. As this is how I have simulated the data, I know

[10] I use the `tree` command of `R`'s library `tree`. It uses the Gini coefficient as the impurity measure, and cross-validation to find the tree depth with the optimal tradeoff between bias and variance.

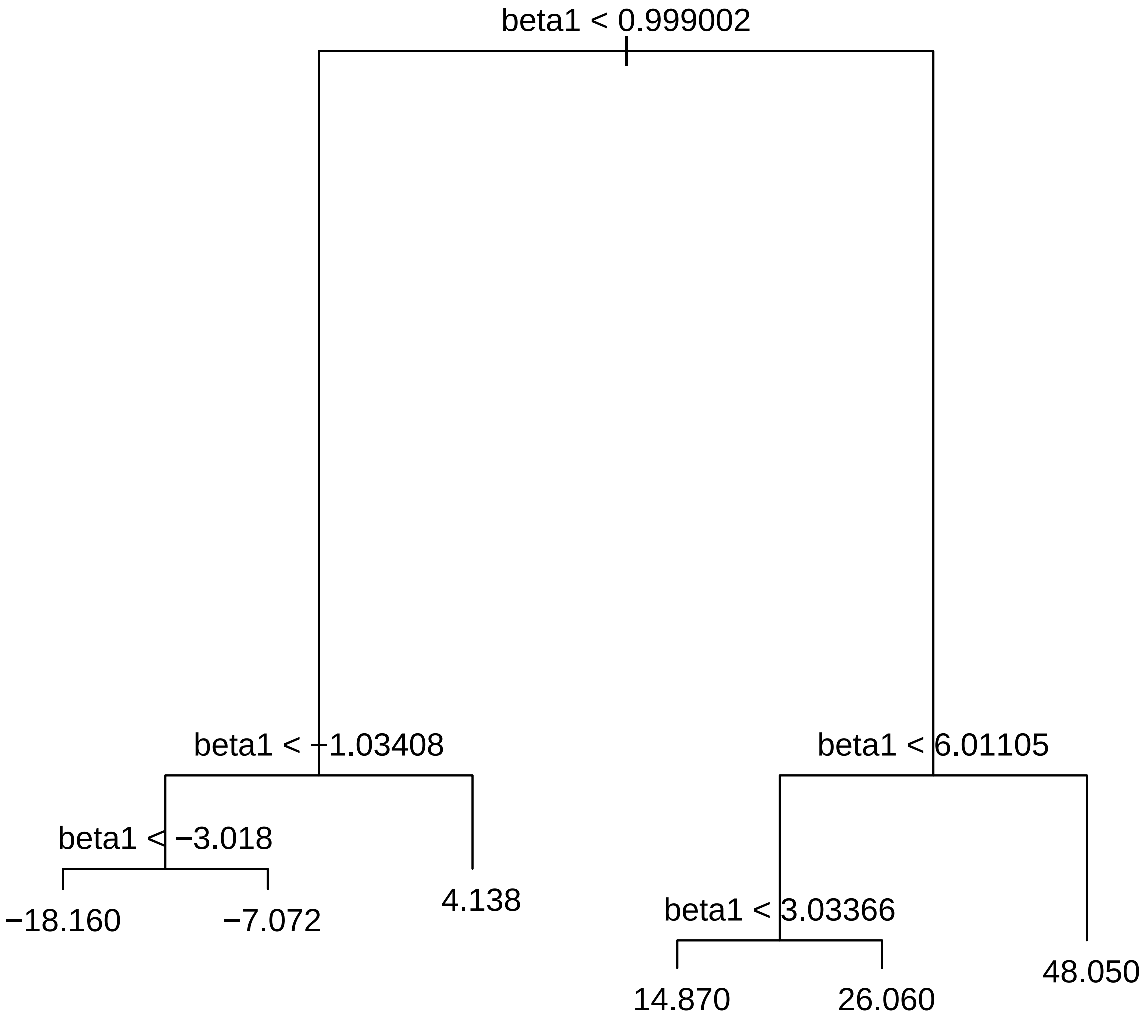

Figure 3: Regression Tree from Scores of Local Regressions

that this will allow me to find the generated types. In a real experiment, it would of course depend on background knowledge whether this assumption seems well founded.

On this assumption, I can use Wald tests to estimate the effect of treatment conditional on type. As CART (together with the observed treatment classifier) finds 9 types, one may first want to know the estimated overall treatment effect, which is $.225, p < .001$. $\hat{\tau}_1$ and $\hat{\tau}_2$ correspond to type 1 in Figure 2. The time trend is strong when untreated, and approximately twice as strong when treated. The treatment effect is directly measured by $t * \hat{\tau}_2$ in Table 4. CART splits the treated cases corresponding to type 2 in Figure 2 into a few cases (8 participants) of $\hat{\tau}_4$ and a large group (50 participants) of $\hat{\tau}_5$. Wald tests show that the difference to the corresponding untreated type $\hat{\tau}_3$ is in either comparison highly significant ($p < .001$), but the comparison with $\hat{\tau}_4$ is -1.5 and hence has the wrong sign. By contrast, the comparison with $\hat{\tau}_5$ is 2.313 and as expected. $\hat{\tau}_6$ and $\hat{\tau}_7$ correspond to type 3 in Figure 2. By design, for this type treatment is immaterial, which is also what the regression finds: the difference between both interaction effects is $.004, p = .894$. $\hat{\tau}_8$ and $\hat{\tau}_9$ correspond to type 4 in Figure 2. The estimated treatment effect is $-1.978, p < .001$, and hence very close to the data-generating process.

Figure 4 shows that, overall, the local regression approach predicts the data very well. The predicted values from (7) not only reconstruct the six arrows from Figure 1. With one exception, the predicted value even sits close to the midpoint of the local distribution of y. Only for the third arrow from above, the predicted values are at the lower bound of the local distributions. The root mean squared error is less good than for the (technically incorrect) finite mixture model, but with 2.219 still very good.

5 Experimental Data

In the final step, I use the seminal contribution of Fischbacher et al. (2001); Fischbacher and Gächter (2010) to explore the power of the approach with real experimental data. Fischbacher & Gächter have participants play a standard linear public good, where payoff is defined by (8)

$$\pi_i = 20 - c_i + .4 \sum_{k=1}^{4} c_k \tag{8}$$

In (8) π_i is payoff, c_i is the contribution a participant makes to the public good of a group of size $K = 4$. As $.4 < 1$, it is individually rational to keep the endowment. Yet as $4 * .4 = 1.6 > 1$ it is socially rational that all group members contribute their entire endowments. The novelty is the use the strategy method (Selten, 1965). Each participant makes two contribution choices: one unconditional, and one conditional on the mean choice of the remaining participants. After the game, the one group member is randomly determined for whom the conditional choice is payoff relevant.

t	4.005*** (.020)
$\hat{\tau}_2$	.180 (.408)
$\hat{\tau}_3$	-.095 (.390)
$\hat{\tau}_4$	-.989 (.776)
$\hat{\tau}_5$	.337 (.408)
$\hat{\tau}_6$	.344 (.432)
$\hat{\tau}_7$	.510 (.427)
$\hat{\tau}_8$	.029 (.408)
$\hat{\tau}_9$	-.178 (.408)
$t * \hat{\tau}_2$	3.978*** (.028)
$t * \hat{\tau}_3$	-2.326*** (.027)
$t * \hat{\tau}_4$	-3.846*** (.053)
$t * \hat{\tau}_5$	-.033 (.028)
$t * \hat{\tau}_6$	-4.049*** (.030)
$t * \hat{\tau}_7$	-4.045*** (.029)
$t * \hat{\tau}_8$	-6.016*** (.028)
$t * \hat{\tau}_9$	-7.994*** (.028)
cons	3.956*** (.288)
N uid	400
N obs	4000

Linear random effects model, based on estimated types. Standard errors in parenthesis. *p<0.05; **p<0.01; ***p<0.001

Table 4: Two-Step Approach: Final Model

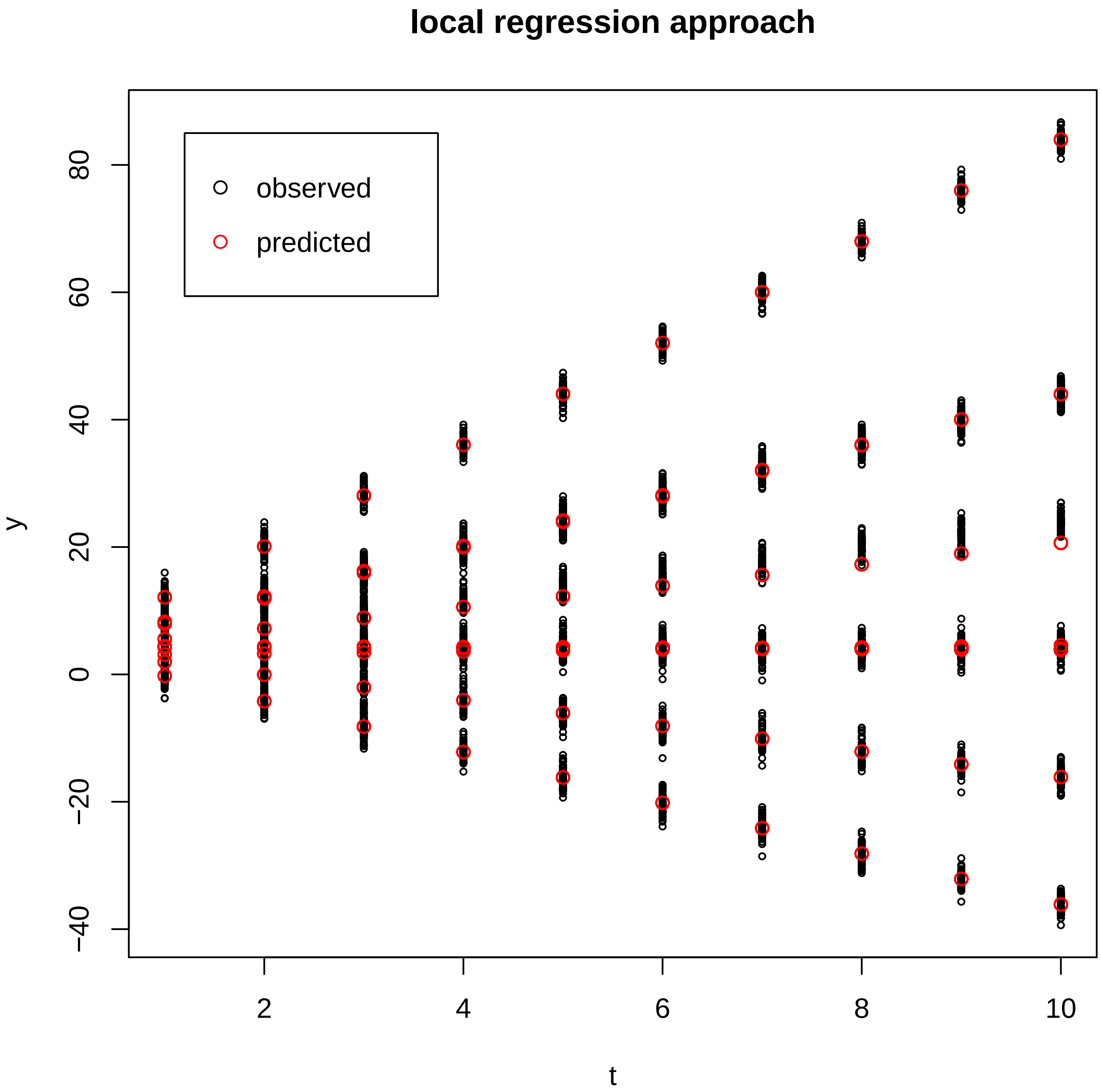

Figure 4: Raw Data and Predicted Values from Local Regression Approach

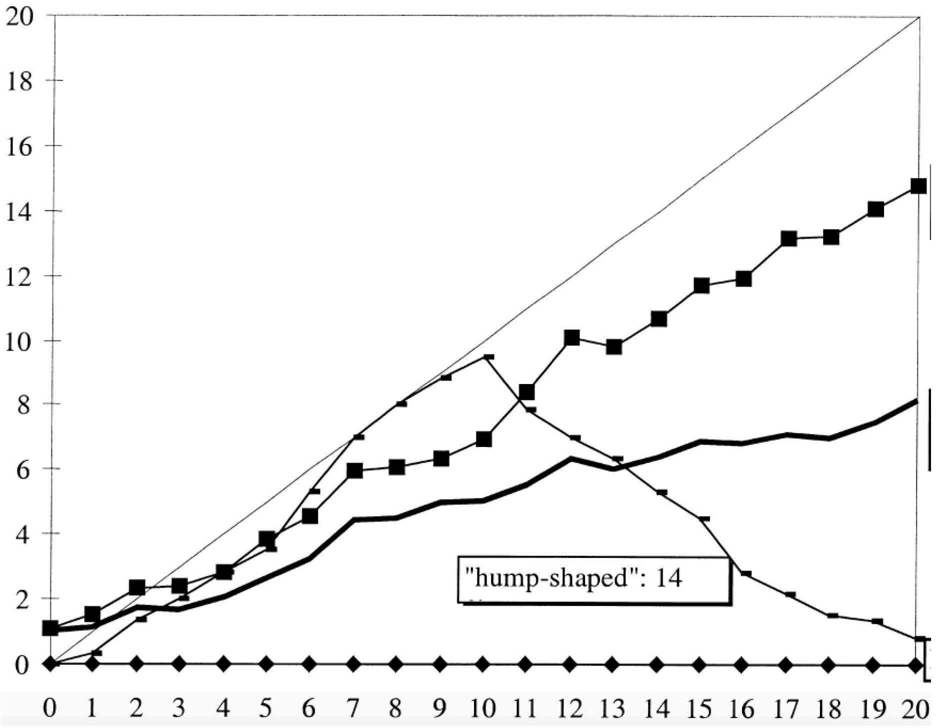

Figure 5: Cooperation Types, Fischbacher Gächter Economics Letter 2001, Fig.1: x-axis c_{-i}, y-axis c_i

For this participant, the design removes strategic uncertainty. This provides a clean test of conditional cooperation: if, but only if, others are holding back the pull of selfishness, conditionally cooperative participants are happy to do so as well. This is indeed what Fischbacher & Gächter find for 50% of their participants. Yet 30% free ride, and 14% exhibit a peculiar pattern of behaviour: as long as the contributions of others are moderate, they match them. But if others contribute more than half of their endowment to the public project, they exploit them, the more so the more they contribute, see Figure 5.

In their original, frequently cited contribution, Fischbacher & Gächter had only 44 participants. In a later paper, they have repeated the test with a larger sample of 140 participants, and have made the data available (Fischbacher and Gächter, 2010). I apply my proposed method of organizing the type space to this new dataset.

The research question can be formulated in statistical terms as (9)

$$c_i = \beta_0 + \beta_1 c_{-i} + \epsilon_i + \epsilon_{li} \tag{9}$$

Each participant makes $L = 21$ conditional choices c_i, for the case that the remaining three group members on average unconditionally contribute $l = c_{-i} = 0..20$ tokens. As the participant in question stays the same, a specification is in order that filters out

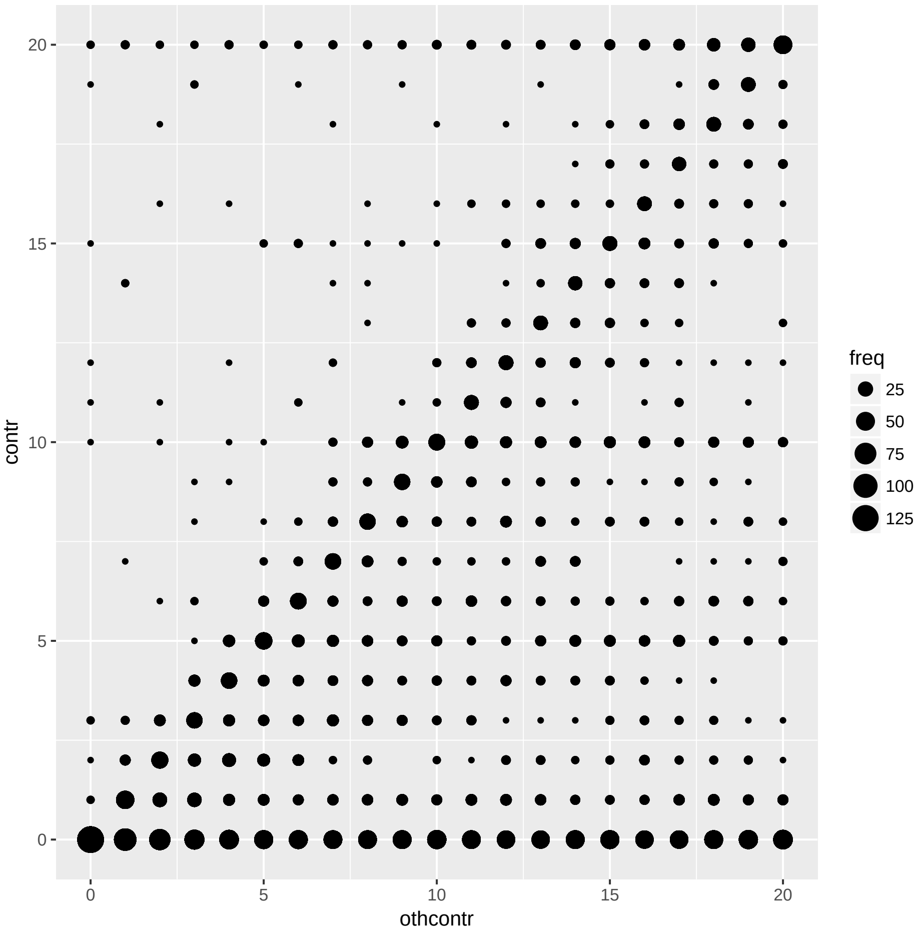

Figure 6: Fischbacher & Gächter Distribution of Raw Data.
Note that the graph represents choices, not individual choice patterns

unobserved individual idiosyncrasies with the random effect ϵ_i. If one estimates (9), one finds $\beta_0 = .531(p = .201)$ and $\beta_1 = .425(p < .001)$. This naive model thus suggests a population of imperfect ($\beta_1 < 1$) conditional cooperators. Yet Figure 6 shows clear heterogeneity. Inspecting the figure suggests three relatively clear types: those who freeride and always contribute 0; those who unconditionally believe in the common good and always contribute 20; and those who perfectly condition their own choices on the mean choice of their group members and exhibit choices on the 45° line. Yet many choices do not match either of the three patterns. Such choices are more frequent below than above the 45° line. Attempts at estimating a finite mixture model fail, even if I only impose 2 or 3 types.

I instead use my proposed method to organize the type space. In this experiment, treatment is exclusively within subjects. It consists of the contribution c_{-i} on which the respective participant i is allowed to condition her contributions. Hence there is no need to use steps 9 and 11-12 of the Algorithm.

	1	2	3	4	5	6	7
1	44	1	0	0	0	0	0
2	0	0	0	0	3	3	0
3	13	4	0	0	3	0	1
4	0	0	0	0	2	3	3
5	0	0	3	0	0	0	0
6	0	0	0	3	0	0	0
7	1	6	0	0	6	0	1
8	0	0	0	0	4	1	35

horizontal axis: types based on local regressions with only a linear term; vertical axis: types based on local regressions with a linear and a quadratic term. Numbers are frequencies.

Table 5: Type Space

Figure 7 collects the results. The upper left panel is resulting from, separately for each participant, regressing c_i on c_{-i}. As Figures 5 and 6 suggest the possibility of a non-linear relationship, the upper right panel of Figure 5 is derived from local regressions of c_i on $c_{-i}+c^2_{-i}$. The former exercise yields 7 distinct types, the latter 8. As Table 5 shows, both methods agree for the extreme cases (types 1, and types 7 linear vs. 8 quadratic), but disagree in the intermediate range. Both trees agree that the slope of the individual reaction curve (β_1 in either local regression) is most important, and hence defines the first split. Yet the tree based on linear models already splits at moderate inclination to condition on c_{-i} ($\beta_1 = .356$), while the tree based on quadratic models requires $\beta_1 = .782$. The intermediate range ($.356 < \beta_1 < .711$) is assigned to a separate type in the tree based on linear models. For this tree, all finer grained separation is based on the intercept of local regressions. By contrast, the tree based on quadratic models uses the coefficient of the quadratic term β_2 in the local regressions for classification in either branch of the tree (for details see Figure 8). There is no statistical reason to prefer one approach over the other. The choice should depend on the conviction of the researcher about the importance of non-linearities in the reaction function.

The most instructive graph is, however, Figure 7. For each type, it aggregates over conditional choices, separately for each possible (mean) unconditional choice. Whether local regressions include a quadratic term or not (upper right and upper left panels), there is a type that almost perfectly matches the unconditional choices; a type that is almost perfectly selfish; a type with very high contributions even if the unconditional contributions are low. Characteristics of the types in the middle differ. If one includes the quadratic term in the local regressions, there is a type that imperfectly matches the

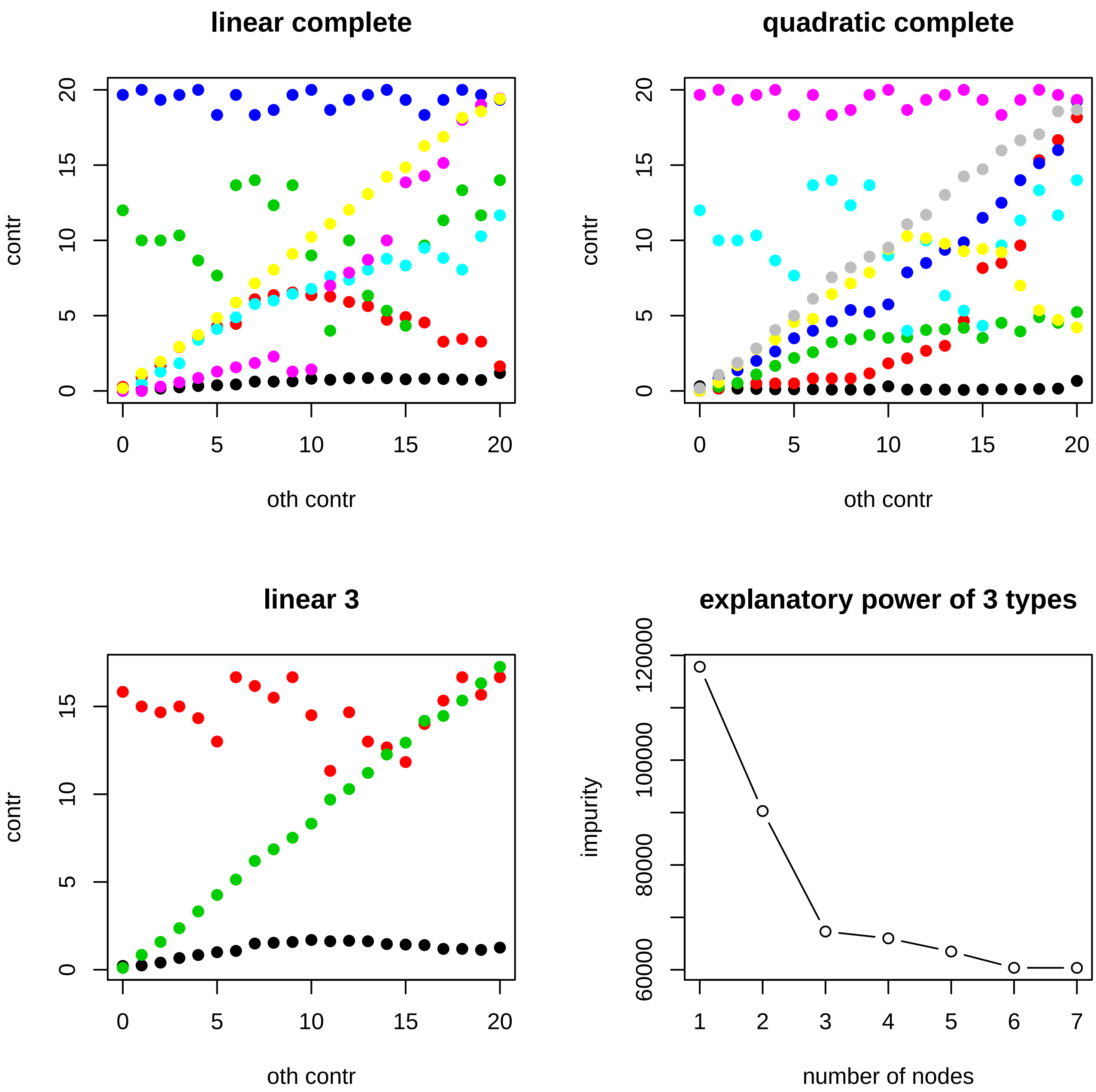

Figure 7: Types Induced by CART

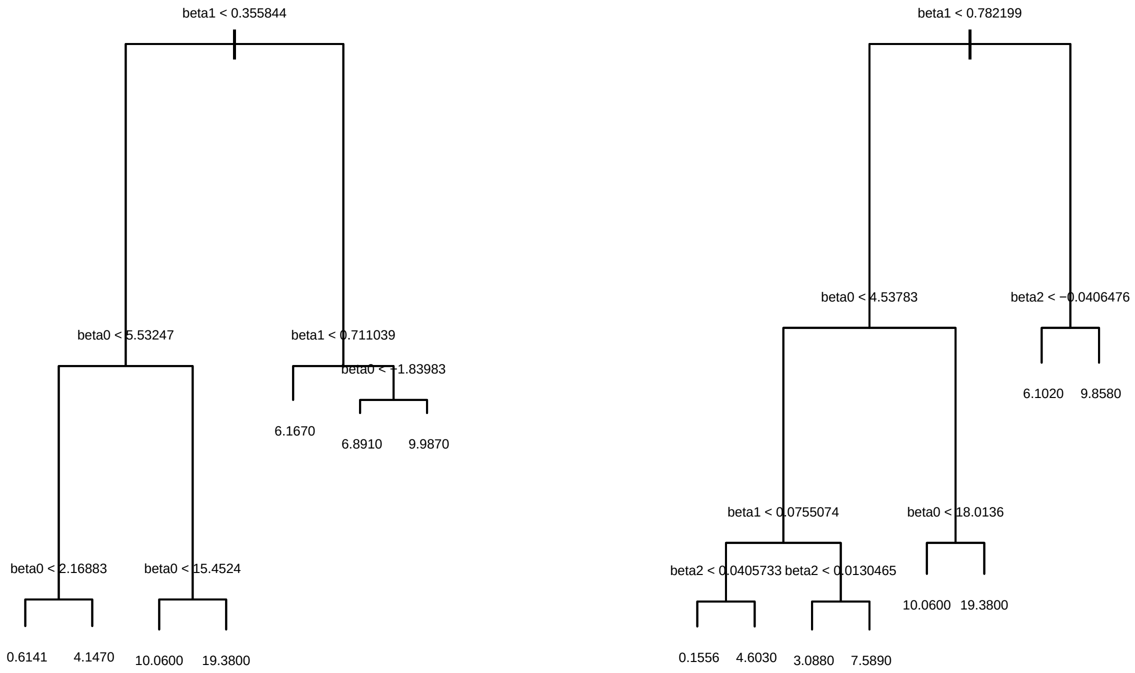

Figure 8: Trees Induced by Local Regression with Only Linear / also Quadratic Term

unconditional choices; a type that matches very low unconditional choices, but then levels off; a type that is selfish if unconditional choices are low, but comes closer with higher unconditional choices.

One can make sense of all these types. The optimality criterion of CART suggests that one would not run an excessive risk of overfitting. Yet the lower right panel of Figure 7 shows that the loss in precision is low if one only allows for three types. The choice patterns of these three types are shown in the lower left panel. The largest type (69 participants) is actually the (almost) selfish type. The type that almost perfectly matches the unconditional choices is a little less frequent (65 participants). There is finally a small type (6 participants) that makes high contributions, irrespective of the contributions made by the remaining group members. Note that this partition of the type space results from a "pruned" tree, with just three final nodes. Hence it assigns all participants to a type, not only those who exhibit patterns similar to the ones shown in the lower left panel when allowing for 7 types (upper left panel). This is remarkable as the reduced type space leads to clearly discernible choice patterns, despite the fact that it has to assign all participants to one of these three types.

Hence the proposed method corroborates what is often treated as a stylized fact in the community: the typical experimental community consists of large groups of conditional cooperators and selfish participants, and a small group of altruists.

6 Discussion

The data from economic experiments often suggests patterned heterogeneity. Reactions to treatment do not only vary. They seemingly vary in systematic ways. In the long run, one would wish to theorize the type space, and have reliable measures for classifying participants into types. But an important first step in the research process is organizing the type space. In principle, estimating heterogeneous treatment effects is a job for a finite mixture model. Such a model simultaneously estimates the probability that a given observation falls into one of the types, and the reaction of participants from this type to treatment. Yet these models have a number of drawbacks: (a) one must posit the number of types, and cannot take them from the data; (b) experimental data is frequently repeated, and often also interactive. Finite mixture models have a hard time capturing the dependence at the individual (and possibly group) level; (c) finite mixture models require two-dimensional maximum likelihood estimation. The datasets from experiments are often too small for these demanding models to converge.

In this paper I propose a simple two-step procedure to address these concerns. This procedure exploits the panel structure of many experimental datasets. Separately for each participant, I estimate a local regression of choices on those variables that change over time. I use the coefficients from these local regressions to train a classification algorithm. Specifically I propose to estimate a regression tree that uses the coefficients from the local regressions to predict choices in the experiment. This procedure allows to assign each participant to a type. If treatment is between subjects, I interact this classification with treatment. I propose to use the standard procedure for regression trees to find the optimal number of types (a). The final step of the procedure can easily handle random and mixed effects models, as at this point type need no longer be estimated (b). And splitting up the definition of the type space, and using type for explaining treatment effects, drastically facilitates estimation, so that in my trials the model always converged (c).

The proposed approach has a number of limitations that are worth spelling out. Local regressions require within participant variation. Hence the method does not work with one-shot experiments. Yet the variation need not result from reaction to treatment. Any variation resulting from repeated reactions suffices. It of course is for the researcher to justify that such variation is meaningful for finding types that exhibit systematically different reactions to treatment.

The researcher must be confident to assume that type is a personality trait, and hence does vary between, but not within participants.

Technically, the approach works as soon as each participant is observed more than once, even if further observations are from supplementary tests, not from the main experiment. Yet the shorter the panel, or the more remote supplementary tests are from the main experiment, the less one will be confident that one precisely captures patterned reactions to treatment in the population.

The approach is straightforward if treatment varies within participants, i.e. in experimental jargon in a within subjects treatment. If treatment exclusively varies between subjects, the approach allows to precisely estimate reactions to treatment conditional on type. One can also precisely estimate the reactions of different untreated types to change over time. Yet without additional information, or suitable assumptions, one cannot match one untreated to one treated type. In the simulated data of Figure 1, one cannot say whether the two upper arrows are of one type (as they indeed are in the simulated data), or whether the uppermost arrow is how one of the other arrows with black dots react to treatment. If it is important for the interpretation of the treatment effect to get this match right, and if the experimenter suspects a heterogeneous reaction to treatment, a hybrid design would be appropriate: one not only tests the treatment effect between, but also within subjects. Then the within component can be used for type classification.

At each step, CART implements the binary split of the data that explains most of the (remaining) variance. If one draws random samples from a larger population, the trees tend to exhibit some variance. If one is concerned about this possibility, one can use multiple starting points or bootstrapping (which the machine learning community calls bagging). The coefficients from local regressions are usually not hugely different from each other. The more they are, the more it would be likely that the coefficients with higher variance have a higher impact on the resulting tree. If one is concerned about this, one can standardize the coefficients before building the tree. Finally, if one coefficient exhibits higher variance than another, it likely will receive greater importance in organizing the type space. For this application, this effect tends to be desirable. But if one were concerned, one could use the procedure that the machine learning community calls boosting. One builds multiple trees, and averages types over these trees. Each tree randomly drops variables from the dataset. Yet if the local regression is simple, as in the examples presented in this paper, boosting would be inappropriate. One would frequently drop the information that should be most important for classification. At any rate, both bootstrapping (bagging) and boosting, i.e. what the machine learning community calls a random forest, will only yield types. One does not have a single, easily interpretable tree.

The local regressions are not meant to predict a population effect. The fact that a coefficient in a local regression is insignificant is therefore not per se a matter of concern. The coefficients are just a way to characterize participants (cross sections). Yet the fact that different participants react in more or less discernible ways to changes over time may induce a different degree of confidence in this characterization. If different participants exhibit very different consistency in their reaction to changes over time, one might want to rely more on the information from participants whose reactions can be estimated more precisely. Weighted estimation is not standard for CART. Yet one can emulate weighting by the inverse of precision by multiplying the data, and adding the more (identical) datapoints the more the individual estimate is precise. For detail, please see the Appendix.

Arguably, many behavioral traits are not universal. These traits are also not just more or less pronounced. There are discernible types. Yet organizing the type space is challenging. This paper proposes a simple and robust method to do so, provided the experiment is repeated.

References

Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Nicholas Bardsley and Peter G Moffatt. The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, 62(2):161–193, 2007.

Leonardo Becchetti, Vittorio Pelligra, and Francesco Salustri. Testing for heterogeneity of preferences in randomized experiments: a satisfaction-based approach applied to multiplayer prisoners dilemmas. *Applied Economics Letters*, 24(10):722–726, 2017.

Paolo Berta, Salvatore Ingrassia, Antonio Punzo, and Giorgio Vittadini. Multilevel cluster-weighted models for the evaluation of hospitals. *Metron*, 74(3):275–292, 2016.

Marco Bertoletti, Nial Friel, and Riccardo Rastelli. Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, 73(2):177–199, 2015.

Friedel Bolle, Yves Breitmoser, and Steffen Schlächter. Extortion in the laboratory. *Journal of Economic Behavior & Organization*, 78(3):207–218, 2011.

Marco Bonetti and Richard D Gelber. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481, 2004.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):211–229, 2016.

Antoni Bosch-Domènech, José G Montalvo, Rosemarie Nagel, and Albert Satorra. A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental Economics*, 13(4):461–475, 2010.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. *Wadsworth International Group*, 1984.

Yves Breitmoser. Strategic reasoning in p-beauty contests. *Games and Economic Behavior*, 75(2):555–569, 2012.

Alexander L Brown and Hwagyun Kim. Do individuals have preferences used in macro-finance models? an experimental investigation. *Management Science*, 60(4):939–958, 2013.

Adrian Bruhin, Ernst Fehr, and Daniel Schunk. The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, 2018.

Anna Conte and M Vittoria Levati. Use of data on planned contributions and stated beliefs in the measurement of social preferences. *Theory and Decision*, 76(2):201–223, 2014.

David J Cooper and E Glenn Dutcher. The dynamics of responder behavior in ultimatum games: A meta-study. *Experimental Economics*, 14(4):519–546, 2011.

Partha Deb and Pravin K Trivedi. Finite mixture for panels with fixed effects. *Journal of Econometric Methods*, 2(1):35–51, 2013.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Christoph Engel. Dictator games: A meta study. *Experimental Economics*, 14(4): 583–610, 2011.

Urs Fischbacher and Simon Gächter. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–56, 2010.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001.

M Gail and R Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pages 361–372, 1985.

Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Bachir Kassas, Marco A Palma, and Charles R Hall. Self-serving motivations of high- and low-income individuals in public goods provisions. Technical report, 2018.

Jaromír Kovářík, Friederike Mengel, and José Gabriel Romero. Learning in network games. *Quantitative Economics*, 9(1):85–139, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. Technical report, 2017.

Tony Lancaster. The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413, 2000.

Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

Peter G Moffatt. *Experimetrics: Econometrics for experimental economics*. Macmillan International Higher Education, 2015.

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high-dimensions. Technical report, 2017.

Luís Santos-Pinto, Adrian Bruhin, José Mata, and Thomas Åstebro. Detecting heterogeneous risk attitudes with mixed gambles. *Theory and Decision*, 79(4):573–600, 2015.

Willi Sauerbrei, Patrick Royston, and Karina Zapien. Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches. *Computational Statistics & Data Analysis*, 51(8):4054–4063, 2007.

Reinhard Selten. Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperimentes. Seminar für Mathemat. Wirtschaftsforschung u. Ökonometrie, 1965.

Carolin Strobl, James Malley, and Gerhard Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4):323, 2009.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

T Wendling, K Jung, A Callahan, A Schuler, NH Shah, and B Gallego. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, 2018.

Jennifer Zelmer. Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3):299–310, 2003.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Appendix: CART with precision weight

If one wants to weigh datapoints by the precision of estimates from the local regressions, this can be achieved with the following modifications of the algorithm:

Algorithm

```
1. Let $D_0$ be a panel with dependent variable $y_{it}$, and explanatory variables $\boldsymbol{x}_{it}$ that include treatment $\theta_i$ (which may differ over repetitions, i.e. may be $\theta_{it}$)
2. initialize $\boldsymbol{\beta}$ and $\boldsymbol{tval}$ for the t-values of the local regressions
   For every participant Do
3. regress $y_{it}$ on all time varying $\boldsymbol{x}_{it}$
4. collect participant $id$ and all $\boldsymbol{\beta}_i$ as well as $\boldsymbol{tval}$ in separate data frame $D_1$
   EndFor
5. for each datapoint, calculate mean t-value (over all coefficients that feature in the local regression)
6. use critical t-values (taking # df into account) to assign weight to each datapoint (e.g. 5 if p < .001, 4 if p < .01, 3 if p < .05, 2 if p < .1, 1 if p > .1)
7. expand datapoints in $D_1$ by weight (hence add 4 identical datapoints if weight is 5, and none if weight is 1)
8. merge $D_1$ with $D_0$ on $id$
9. fit classification tree of $y_{it}$ on $\boldsymbol{\beta}$
10. use standard algorithm to define optimal depth of tree
11. use optimal tree to assign type to each participant
12. estimate panel version of (2)
```

In the simulated dataset, this procedure assigns weight 1 to 108 original datapoints, weight 2 to 5 datapoints, weight 3 to 41 datapoints, weight 4 to 39 datapoints, and weight 5 to 207 datapoints. The resulting classification tree finds very similar cutpoints, but has a different structure, and one final node less, see Figure 9.

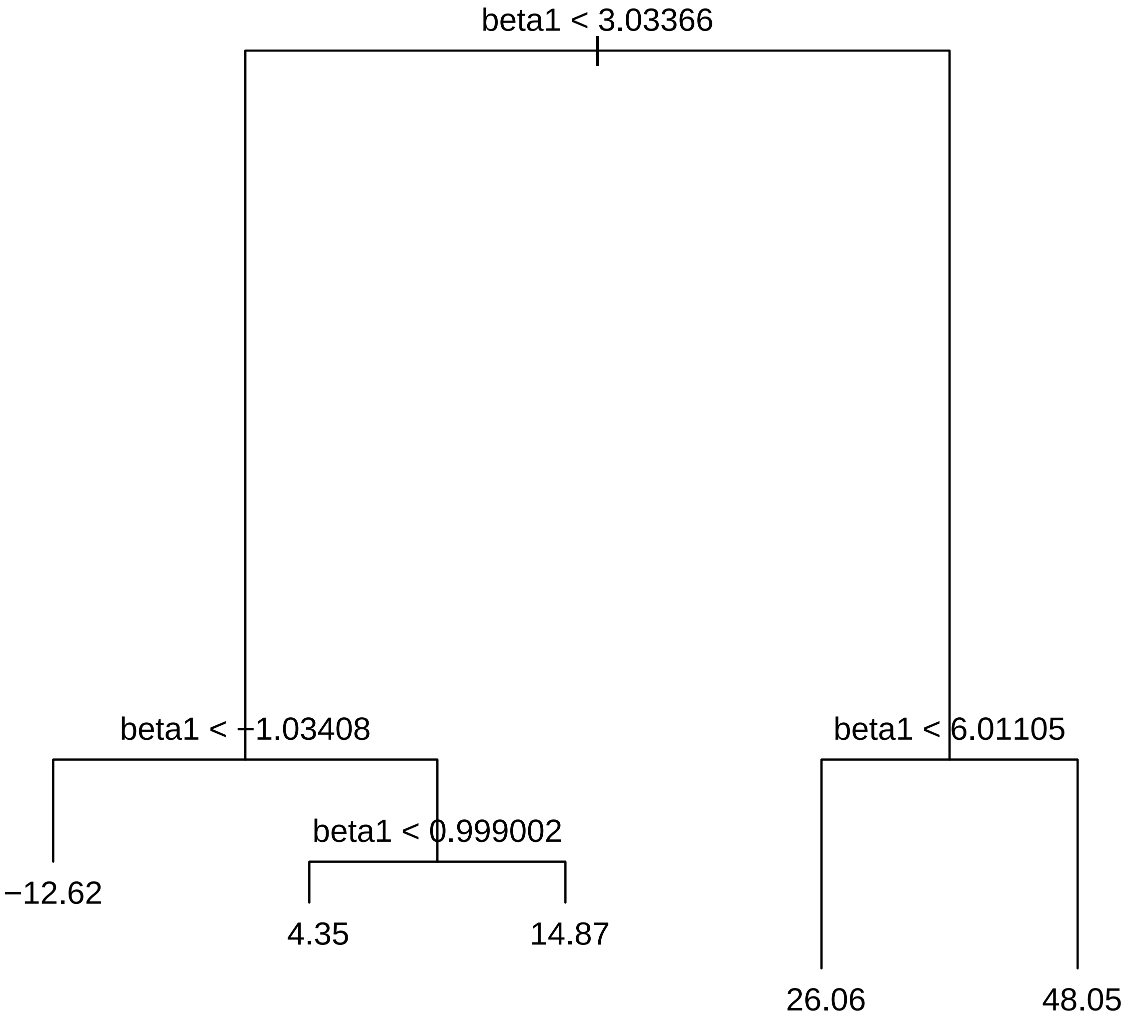

Figure 9: Tree Induced by Precision Weighted Local Regressions