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

This version: January 11, 2020

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 11, 2020

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 finite mixture models may need more data than the experiment provides. The approach in principle requires ex ante knowledge about the number of types. Finite mixture models make distributional assumptions that one may not feel comfortable with. They 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, that does not make distributional assumptions about the type space, 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, and one would be justified to assume that the choice functions of individuals are deterministic, 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 is typically estimated with full information maximum likelihood. One 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. While this approach is appealing, it has a number of drawbacks. If the model is estimated with maximum likelihood, one must

[1] For detail see below Section 6.

make two-dimensional distributional assumptions (like joint normality). These assumptions are not necessarily easy to defend, especially if one expects the number of types to be small, and the distribution of types to be pronouncedly asymmetric. Typical experimental datasets are rather small. As estimation is in two dimensions, finite mixture models with experimental data need not converge. Chances for convergence improve if one tries out alternative statistical models, alternative distributional assumptions, a whole grid of starting values, alternative approximation algorithms or the rescaling of parameters.[2] But one may then be afraid of being on a fishing expedition that yields a false positive result.

Many experiments are repeated, and often additionally interactive. Such an experiment generates data from choices, nested in individuals, nested in groups. Each group is a single independent observation. To properly capture the dependence structure and the heterogeneity of types, one would need four-dimensional maximum likelihood. Challenges for the amount and the purity of the data compound. A further drawback is the necessity to fix the number of types beforehand. Admittedly, there is a way out: in a first step, one estimates a series of finite mixture models, and imposes an increasing number of types. In the second step one uses a technique for the comparison of non-nested models to select the most convincing model (El-Gamal and Grether, 1995; Bosch-Domènech et al., 2010; Bruhin et al., 2018). Yet this process is laborious and there is no consensus about the criterion for model selection.

In this paper, I propose an alternative approach. It is not a panacea. It comes with its own limitations, which I discuss in the concluding sections. Yet it has its comparative advantages. Experimentalists interested in heterogeneous reactions to treatment have another tool in their box. They may revert to it as it tends to be more robust to implement, or as they are sufficiently concerned about one of the limitations of a finite mixture model.

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 (per participant) regressions characterize the individual's type. Standard machine learning techniques 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. The proposed approach is thus a two-steps estimator. Such estimators have been used frequently in other areas of statistics, and in particular for the combination of non-parametric with parametric methods (for an overview see Chapter 17 Greene, 2003).

If treatment is within-subjects, this procedure directly produces estimates of the treatment effect conditional on type. If treatment is (exclusively) between-subjects, there is

[2] I am grateful to an anonymous referee for listing all these options.

an additional challenge. The procedure separates types *conditional* on treatment. Yet if one, for instance, finds 8 conditional types, one does not yet know which untreated and which treated conditional types originate from the same unconditional type. One unconditional type may be very sensitive to treatment, while another is not. One unconditional type may even react positively to treatment, while another reacts negatively. Hence one needs additional assumptions (or complementary within-subjects data) to match types of untreated with types of treated subjects. Yet 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.

Machine learning is not standard in the experimental literature. A few remarks explaining the approach may therefore be in order (for an excellent introduction see James et al., 2013). The epistemic goal of machine learning is not identification, but prediction or, as exploited here, classification. One isolates the observed features that have the highest potential to define discernible classes in the dependent variable of interest. This paper uses one of the most well-established methods in machine learning, a classification and regression tree CART. The method is conceptually very simple. One starts by finding the independent variable that most clearly separates the dependent variable into two classes. This is the origin of the tree. Now one repeats the procedure separately for each branch of the tree, until the final tree is constructed[3]. As the simulations in sections 4 and 5 show, this very simple approach is very powerful. In the conclusions, I discuss in which situations more elaborate machine learning procedures might be of interest, and explain their logic.

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 investigates the robustness of the method. Section 6 applies the approach to a real experimental dataset. Section 7 concludes with discussion. The code for implementing the procedure in R is reproduced in the technical appendix.

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 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

[3] For more about the method, and in particular the definition of the depth of the tree, see below section 4.

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 from participants $i \in \{1, ..., N\}$ with a set of dummy variables $k \in \{1, ..., K\}$ indicating type and the indicator function $\mathbb{1}$ assigning type to the individual in question, and with treatment $\theta_i \in \{0, 1\}$. Hence one estimates

$$y_i = \beta_0 + \beta_1\theta_i + \sum_{k=2}^{K} \beta_k \mathbb{1}(k_i = k) + \sum_{k=2}^{K} \beta_{K-1+k} \mathbb{1}(k_i = k) \cdot \theta_i + \epsilon_i. \tag{1}$$

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 $\beta_0 + \beta_k$, and the choice when treated is estimated by $\beta_0 + \beta_1 + \beta_\tau + \beta_{K-1+k}$. In this specification $\beta_2, ..., \beta_K$ 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.[4]

Finite Mixture Model

If type k_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

$$f(y_i) = \sum_{k=1}^{K} \pi_k f_k(y_i | \boldsymbol{x_i}'\boldsymbol{\beta}). \tag{2}$$

In (2) $\boldsymbol{x_i}'\boldsymbol{\beta}$ is a generic way of writing the first two terms in (1), while allowing $\boldsymbol{x_i}$ to contain further covariates, together with the treatment variable θ_i. Yet through π_k, 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 $\sum_k^K \pi_k = 1$.

The model defined in (2) can be estimated with (full information) maximum likelihood. The probabilities (often also referred to as mixing proportions) $\pi_1, ..., \pi_K$ result from unobserved types. Estimating these latent types 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.

Two-Steps Estimator

The previous approaches have treated each data point as an independent observation. Economic experiments are frequently repeated, i.e. come from period $t \in \{1, ..., T\}$. For estimating the treatment effect, this is not a concern. One can estimate the random effects model (3):

$$y_{it} = \beta_0 + \beta_1 \theta_{it} + u_i + \epsilon_{it}, \tag{3}$$

assuming that individuals $i \in \{1, ..., N\}$ in each period $t \in \{1, ..., T\}$ are randomly exposed to treatment $\theta \in \{0, 1\}$, and that choices are nested in individuals i. Yet

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

finite mixture models for panel data are difficult, at least if one wishes to estimate a random effects model. The individual specific error u_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 (2) (Deb and Trivedi, 2013).[5] Another option is demeaning. But either way one loses the ability to estimate the effect of time invariant independent variables.

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. Then type can be inferred from development of individual choices over time. Typical repeated experiments fulfil these conditions. A classic illustration is a public good. But the approach can also be justified for repeated market experiments, or for learning experiments. Actually if the researcher does not have ex ante knowledge about the composition of the type space and cannot pre-classify participants, detecting characteristic patterns in their choices over time is the only possibility to partition the type space from the data.

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 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

[5] 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).

as possible. Hence at the first split, CART uses each coefficient in $\boldsymbol{\beta}$. If coefficients are continuous, CART not only tries out each coefficient, but each cutpoint on each coefficient. This first step creates a tree with one node and, consequently, 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, and 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 method is appropriate for partitioning the type space, 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 $\beta_0 < 2$, and a second split for the right branch of the tree at $\beta_1 < 5$. These splits can be used to assign participants to types. All participants with $\beta_0 < 2$ are classified as $\hat{k}_1$. Participants with $\beta_0 \geq 2$ and $\beta_1 < 5$ are classified as $\hat{k}_2$, and participants with $\beta_0 \geq 2$ and $\beta_1 \geq 5$ are classified as $\hat{k}_3$. Note, of course, that types are estimates. Each participant is assigned to the type that she is most likely to adhere to. To make the estimated character apparent, I use the $\hat{k}$ notation.

One uses these estimated types to predict 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 $u_i + \epsilon_{it}$, 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.[6] This dependence structure can be captured by $u_g + u_{gi} + \epsilon_{git}$, 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

1. Let D_0 be a panel with dependent variable y_{it}, and explanatory variables $\boldsymbol{x}_{it}$ that include treatment θ_i (which may differ over repetitions, i.e. may be θ_{it})

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

```
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 (1)
   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
```

The approach combines a non-parametric first step (CART) with parametric estimation, using the types predicted in the first step as explanatory variables in a parametric model (the random or mixed effects model explaining choices with type and treatment). Such two-step estimators are routinely used in other areas, like selection models or semiparametric regression. Coefficients estimated in the second step are consistent, provided the first step yields a consistent estimate of the input into the second step (Greene, 2003, Chapter 17.7). In general, CART consistently estimates the type space (Breiman et al., 1984, Chapter 12), although proofs still seem to be missing in case the tree is trimmed with pruning (Toth and Eltinge, 2011), which is desirable to avoid overfitting (and proposed in step 7 of the Algorithm).

Yet the literature about two-step estimators is usually not concerned about the consistency of the estimates for the coefficients, but about standard errors. If one uses maximum likelihood for the second step, this is limited information maximum likelihood, not full information maximum likelihood. In the second step, one uses a predicted value from the first step, and ignores the noise inherent in estimating the first step. If both steps are parametric, one can correct standard errors in the second step, using the procedure developed by Murphy and Topel (1985). Yet this method needs a variance covariance matrix from the first step, which is unavailable for non-parametric first steps. If the non-parametric method has a close parametric equivalent, one can take the variance covariance matrix from auxiliary estimation of this parametric alternative (Ackerberg et al., 2012). But there is no close parametric equivalent to CART. Yet on closer scrutiny, this is a second order problem. The first order difference between (full information maximum likelihood) finite mixture and the method proposed here is the way how individuals are assigned to types. While a finite mixture model calculates the probability of participants with a certain vector of choices and covariates to be of *any* of the estimated types, CART assigns each participant to one defined type. Hence with a finite mixture model, assignment of type is probabilistic, while it is deterministic with CART.

In the Appendix I propose a safeguard that at least partly addresses the concern. I exploit the fact that the development of choices over time is used to characterize individuals (and have assumed that types are nested in individuals). For each participant, this development may be more or less consistent. The degree of consistency can directly be read off the standard errors of the local regression. Hence for each input into CART I not only know parameters (the coefficients of the local regression), but also have a measure for the confidence in these parameters. I propose an alternative version of the algorithm that weights the information about the individual participant with the inverse of the standard error from the respective local regression. That way the more the information about a participant is precise the greater its impact on the partition of the type space.

4 Simulation

In this section, I show with simulated data how the approach works. The `R` code for performing the analysis is available in the Appendix, 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 ($\theta_i \in \{0, 1\}$),[7] and individuals are of types $k_i \in \{1, ..., 4\}$. Types differ in their reaction to treatment. Specifically, dependent variable y_{it} is generated according to (4)

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

[7] When generating the data using (4), θ_i is not coded as a (zero-one) dummy variable as otherwise all untreated observations would be identical. Yet for facilitating interpretation, in the final dataset, θ_i is recoded as a dummy variable.

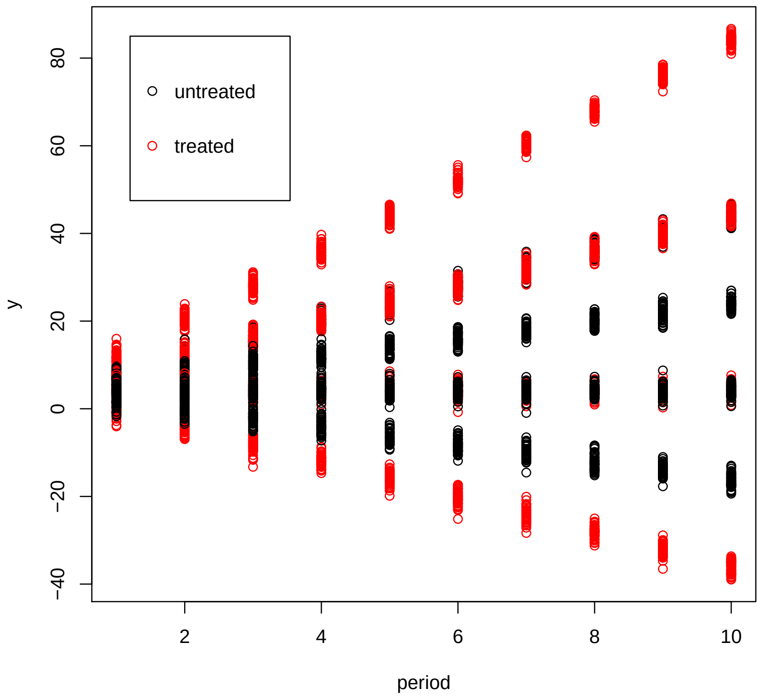

Figure 1: Simulated Data Pooled

where individual specific error $u_i \sim \mathcal{N}(0,1)$ captures dependence within individuals, and $\epsilon_{it} \sim \mathcal{N}(0,1) \perp u_i$ 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 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?[8]

[8] In an experiment, random assignment would exclude 6 types, as there could not be selection of

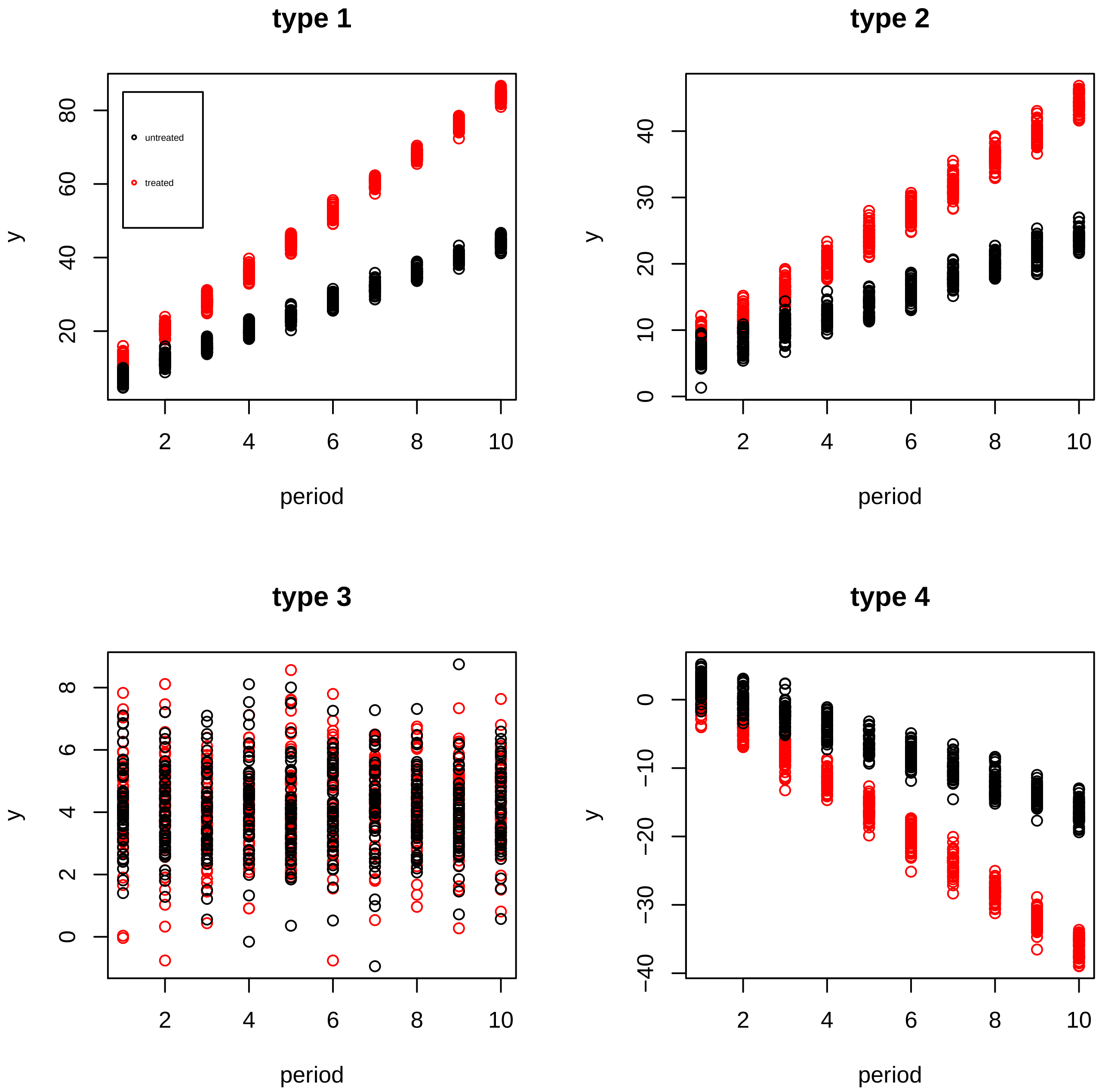

Figure 2: Simulated Data by Type

θ	-0.001 (2.061)
t	0.987*** (0.084)
$\theta * t$	1.009*** (0.118)
cons	4.057** (1.457)
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

As the simulated data generation process is so clean, the correct finite mixture model with four types converges and yields results that closely match (4), Table 2. Yet note that the model ignores dependence at the individual level.[9] As the data is simulated, I can compare estimated with true type. For that purpose I classify an individual to be of the type with the highest posterior probability. The estimate is correct in 96.53 % of all cases. The root mean squared error is 0.0059, which is even considerably less than in the original data, where it is 1.427.[10]

	type 1	type 2	type 3	type 4
p_k	25.25	24.88	24.94	24.93
t	4.013	1.985	-0.003	-1.994
θ	0.315	-0.086	0.105	0.102
$\theta \cdot t$	3.968	2.018	-0.000	-2.006
cons	3.864	4.040	4.101	3.812

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

Table 2: Finite Mixture Model

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

types into treatment.

[9] 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 k and θ. As treatment is exclusively between subjects, the treatment effect would also drop out with demeaning.

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

$$y_t = \beta_0 + \beta_1 t + \epsilon_t. \tag{5}$$

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.[11] 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 four of the types generated by CART exclusively cover treated or untreated cases. Taking this into account, the final set of types consists of eight types, four treated, and four untreated. Actually, CART finds a type space with the exact same frequencies as in the simulated data generating process, i.e. 50 participants per condition.

In the next step I estimate (6), where $\hat{k}_i$ is one of the 8 estimated types.

$$y_{it} = \gamma_0 + \gamma_1 t_{it} + \sum_{k=2}^{8} \gamma_k \mathbb{1}(\hat{k}_i = k) + \sum_{k=2}^{8} \gamma_{2k} \mathbb{1}(\hat{k}_i = k) \cdot t_{it} + u_i + \epsilon_{it}. \tag{6}$$

Table 3 shows that the procedure works well. Type main effects are all insignificant, as they should, given the data generating process of (4) starts at the same point, irrespective of type. The coefficient of t captures the time trend for the first type (it corresponds to type 4 in Figure 2, for the untreated participants). The interaction effects define how much the time trend for each of the remaining estimated types $\hat{k}_i$ differs from the time trend in the first type.

In the simulation, treatment exclusively affects slopes. Consequently the algorithm exclusively uses β_1 for classification. 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 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.

$t \cdot \hat{k}_2$ captures the treatment effect of type 4 in Figure 2. To test the treatment effect for the remaining types, I can use Wald tests. $t \cdot \hat{k}_3$ and $t \cdot \hat{k}_4$ capture the treatment effect for type 3 in Figure 2. As expected, slopes are practically identical (the difference is 0.012), and the difference between the interaction effects is insignificant (p = .591). Untreated and treated cases corresponding to type 2 in Figure 2 are captured in the regression by types $\hat{k}_5$ and $\hat{k}_6$, respectively. The difference in slopes between treated

[11] 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. If users want more flexibility and control in growing and pruning the tree, they can switch to the `rpart` package. There is also a helpful manual for that package (Therneau et al., 2010).

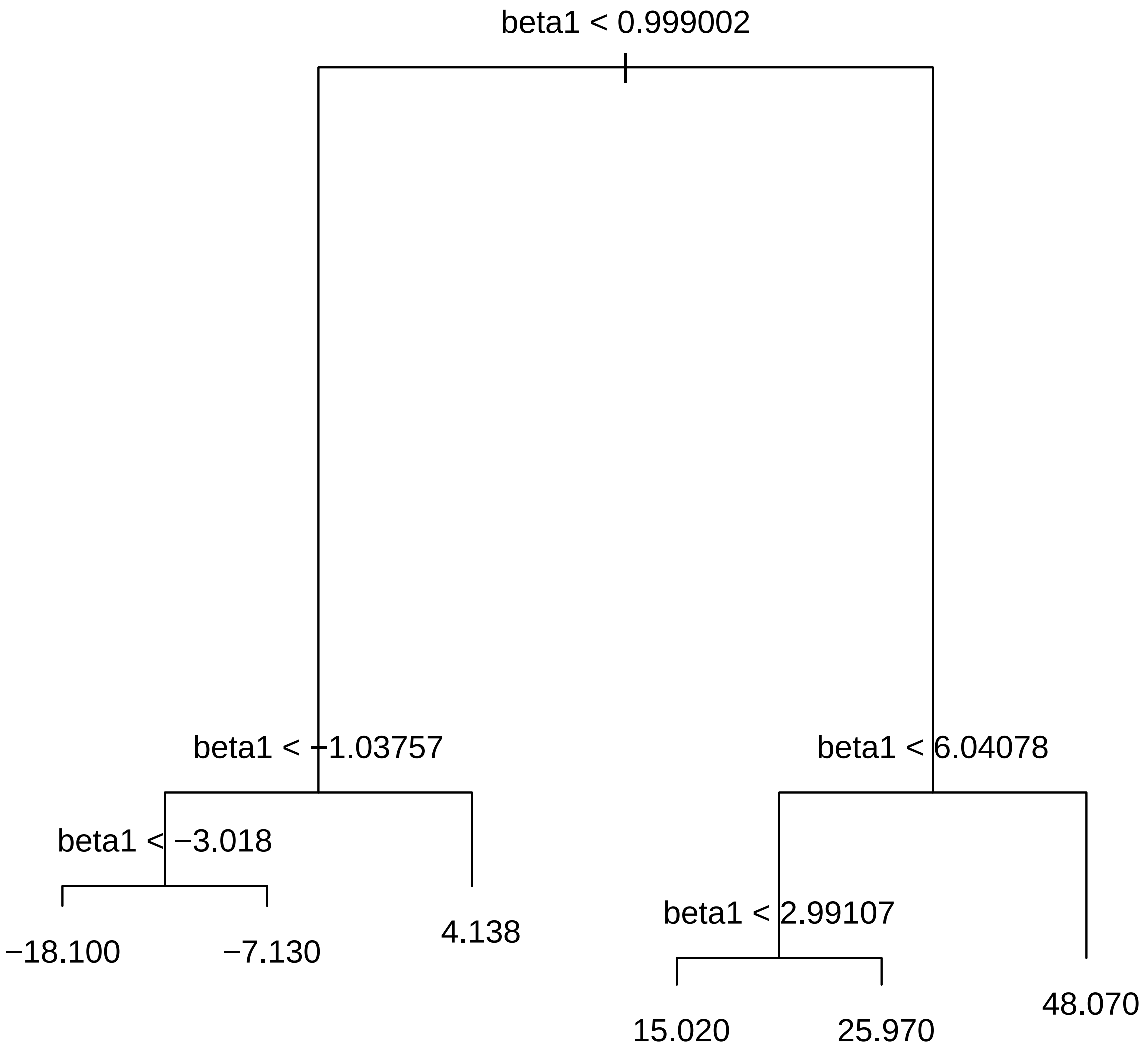

Figure 3: Regression Tree from Scores of Local Regressions

and untreated cases of this type is estimated to be $2.048, p < .001$. Finally untreated and treated cases of type 1 in Figure 2 are captured in the regression by types $\hat{k}_7$ and $\hat{k}_8$, respectively. For this type, the treatment effect is estimated to be $3.991, p < .001$.

t	-2.005*** (0.016)
$\hat{k}_2$	-0.036 (0.247)
$\hat{k}_3$	0.248 (0.247)
$\hat{k}_4$	0.261 (0.247)
$\hat{k}_5$	0.355 (0.247)
$\hat{k}_6$	0.026 (0.247)
$\hat{k}_7$	0.132 (0.247)
$\hat{k}_8$	0.266 (0.247)
$t \cdot \hat{k}_2$	-1.988*** (0.022)
$t \cdot \hat{k}_3$	1.997*** (0.022)
$t \cdot \hat{k}_4$	2.008*** (0.022)
$t \cdot \hat{k}_5$	3.963*** (0.022)
$t \cdot \hat{k}_6$	6.012*** (0.022)
$t \cdot \hat{k}_7$	5.998*** (0.022)
$t \cdot \hat{k}_8$	9.988*** (0.022)
cons	3.900*** (.175)
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 3: Two-Step Approach: Final Model

Figure 4 shows that the local regression approach predicts the data very well. The predicted values from (6) not only reconstruct the six arrows from Figure 1. The predicted values even sit close to the midpoint of the local distribution of y. The root mean squared error of 1.427 is almost perfectly identical with the root mean squared error of the simulated data, which is 1.429.

5 Performance and Robustness

The previous section demonstrates the logic of the approach, and shows that it works well with one simulated dataset. Yet this dataset uses one specific set of random

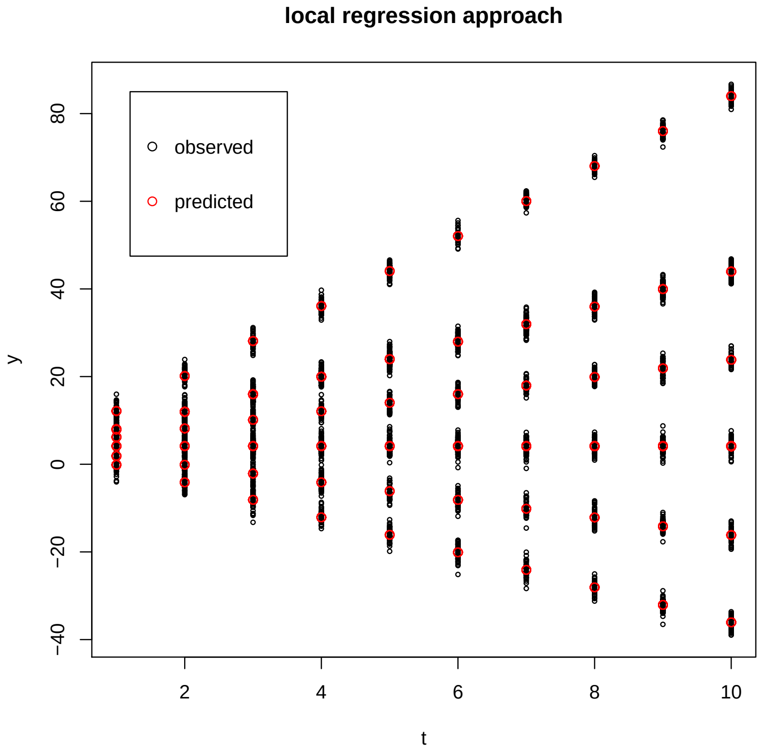

Figure 4: Raw Data and Predicted Values from Two-Steps Approach

variables,[12] and for expositional reasons it works with a very clean data generating process. This section investigates the performance of the proposed two-step estimator, both in comparison with the simulated data generating process and with the alternative to use a finite mixture model. Specifically I rerun the original data generated process 500 times, and I draw 500 samples each of 11 potentially more challenging data generating processes.

In these robustness checks, I alter the data generating process in the ways defined in Table 4:

1. **panel length**: In the original data generating process, the panel is reasonably long ($T = 10$). If the panel is shorter ($T = 5$ or $T = 3$), there is less scope for inferring types from the development of choices over time. For that reason type information is more noisy.

2. **noise**: The main simulation fixes $\sigma_u = 1, \sigma_\epsilon = 1$. In the next step I test whether the two-step estimator still works well with either $\sigma_u = 2$ or $\sigma_\epsilon = 2$.

3. **random coefficients**: In the main simulation, there is individual-specific error u_i, on top of residual error ϵ_{it}, but the individual specific error affects levels of choices, not slopes, while types affect slopes, not levels. I change this by an additional noise term η_i for slopes with $\sigma_\eta \in \{1/10, 1/5\}$.

4. **distribution of the error term**: Finite mixture models are estimated with maximum likelihood. The researcher must therefore specify the distribution of error terms, and of the distribution of types in the population in particular. By contrast the two-step approach finds the type space with CART and must not make distributional assumptions. I investigate in which ways the misspecification of the distribution of individual specific error u_i affects the comparative performance of either approach, by replacing normality with a uniform distribution in the interval $[-2, 2]$, or by a skewed β-distribution with shape parameters 10 and 2.

5. **kinked DGP**: One reason why choices change over time is learning. Now one plausible type of learning stops if individuals believe they have understood the task. I capture this possibility by either letting the more moderate type 2 of Figure 2 or the extreme type 1 stick to the choices they have individually made in period 5 for periods 6-10.

6. **misspecification of dynamics**: The two-step approach critically relies on the information about types contained in the intercept and the slope of the development of choices over time, separately for each individual. In the final simulation I investigate how the approach fares if the data generating process is misspecified.

[12] By fixing the `seed` in `R` at 1234.

I do so by introducing a quadratic term and assume that it is undetected when analysing the data.

A desirable property of an estimator is prediction accuracy. As Figure 5 shows, in this respect the two-step estimator clearly outperforms the finite mixture model. Irrespective of the definition of the data generating process, the two-step estimator is nearly perfect (the mean prediction error is nowhere larger than 0.000000000001). By contrast in many random samples, the prediction error of the finite mixture model is substantial.

Interestingly, the prediction error in an FMM is the higher the longer the panel: with $T = 10$, the second mode of the distribution is near -.5, while it is near -.35 with $T = 5$, and near -.25 with $T = 3$.[13] This is in line with the observation that the finite mixture model ignores the dependence at the individual level. The longer the panel the more important this omission. Increasing u_i or ϵ_i has virtually no effect on the accuracy of the finite mixture model: the density curves for these treatments are almost perfectly masked by the density curve for the original data. Random slopes have opposing effects on accuracy. The second mode of the distribution shifts closer to 0. This reduces the prediction error. Yet the first mode shifts away from 0. This increases prediction error. The latter effect is much more pronounced with $\eta_i = 1/5$. Interestingly, prediction accuracy is higher, not lower, if $u_i \sim \mathcal{U}[-2, 2]$ or $u_i \sim \mathcal{B}(10, 2)$. The fact that, in estimating the finite mixture model, error has been assumed to be normal does not hurt. The fact that choices of one type are kinked does also not reduce, but even slightly increase prediction accuracy. Whether the kink affects a moderate or an extreme type is immaterial: the density curves for both simulations are practically identical and therefore superimposed. Finally overlooking a quadratic term in the data generating process does even lead to the most accurate predictions. Note, however, that even in this best performing case a substantial fraction of the predictions are inaccurate. The inaccuracy is only smaller than in other specifications of the data generating process.

An alternative performance criterion is the fraction of types that have been wrongly classified. As Figure 6 shows, for most specifications of the data generating process, the two-step estimator also outperforms the finite mixture model in this respect. While the fraction of misclassified types is minuscule when the two-step approach is used, it is substantial in the corresponding finite mixture model. Yet with some data generating processes, the two-step model also has problems. The problem is most pronounced if the panel is short ($T = 3$, mean fraction of wrongly classified types 10.68%, maximum 22.75%). The problem is similar with pronounced random slopes ($\eta_i = 1/5$, mean fraction of wrongly classified types 8.48%, maximum 17%), and it remains discernible if a quadratic term in the time trend is ignored (mean 0.64%, maximum 12.75%).[14]

[13] The fact that mean prediction error is much more likely to be negative than positive follows from the asymmetry in the data generating process. There are two types with a positive slope, one of which is pronounced, and only one type with a moderately pronounced negative slope.

[14] The graph is capped at density 20, as the fraction of misclassified types is 0, or close to 0, for most

		T	σ_u	σ_ϵ	σ_η	$\sim$	kink	t^2
1	original	10	1	1	0	$\mathcal{N}$	0	0
2	panel length	**5**	1	1	0	$\mathcal{N}$	0	0
3		**3**	1	1	0	$\mathcal{N}$	0	0
4	noise	10	**2**	1	0	$\mathcal{N}$	0	0
5		10	1	**2**	0	$\mathcal{N}$	0	0
6	random coefficient	10	1	1	$\mathbf{1/10}$	$\mathcal{N}$	0	0
7		10	1	1	$\mathbf{1/5}$	$\mathcal{N}$	0	0
8	distribution	10	NA	1	0	$\mathcal{U}$		
9		10	NA	1	0	β	0	0
10	kinked	10	1	1	0	$\mathcal{N}$	**type 2**	0
11		10	1	1	0	$\mathcal{N}$	**type 1**	0
12	dynamics	10	1	1	0	$\mathcal{N}$	0	**1**

Table 4: Performance and Robustness
500 simulation runs per each of the 12 conditions. T number of periods. σ standard deviation of normally distributed error term with $\mu = 0$, for individual specific error u, residual error ϵ and error term η introducing randomness into slopes. $\sim$ distribution of u as either normal ($\mathcal{N}$), uniform with range $[-2, 2]$ ($\mathcal{U}$) or beta with shape parameters {10,2} (β). Kink for either type 2 or type 1, such that choice in period 5 is kept constant for remaining types. Dynamics: choices develop with $1/20 * t^2$, which is not taken into account in local regressions.

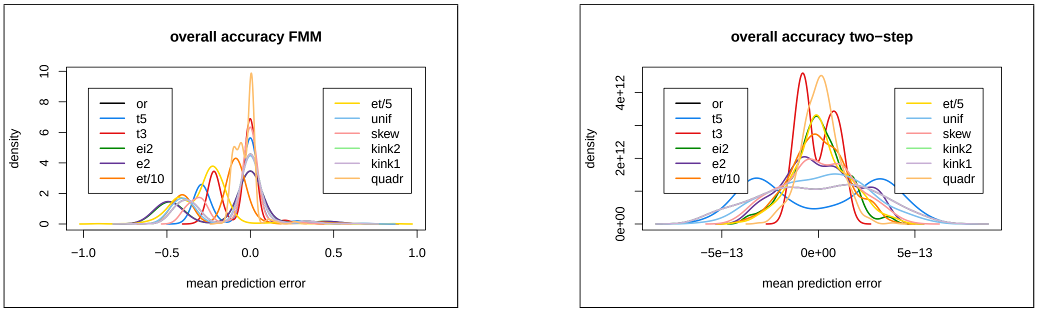

Figure 5: Prediction Accuracy
left panel: finite mixture; right panel: two-steps

In the original data generating process, a finite mixture model gets the type space quite frequently wrong, with modes near a quarter and a third of all cases. This is similar with $\sigma_u = 2$ and $\sigma_\epsilon = 2$: the three density curves are superimposed. While prediction accuracy was better the shorter the panel, the opposite holds for the accurate estimation of the type space. The fraction of wrongly classified types is highest with $T = 3$, lower with $T = 5$, and lowest with $T = 10$. If the randomness in the reaction of individuals to time is more pronounced ($\eta_i = 1/5$), misclassification becomes more frequent than with $\eta_i = 1/10$. Again the fact that maximum likelihood assumes normality does not hurt. Actually the classification of types is even better than in the original data generating process if u_i is taken from a uniform or from a beta distribution. A kink in the reaction of one type to time is also not causing massive misclassification, while ignoring a quadratic term in the reaction time does lead to the most substantial misestimation of the type space.

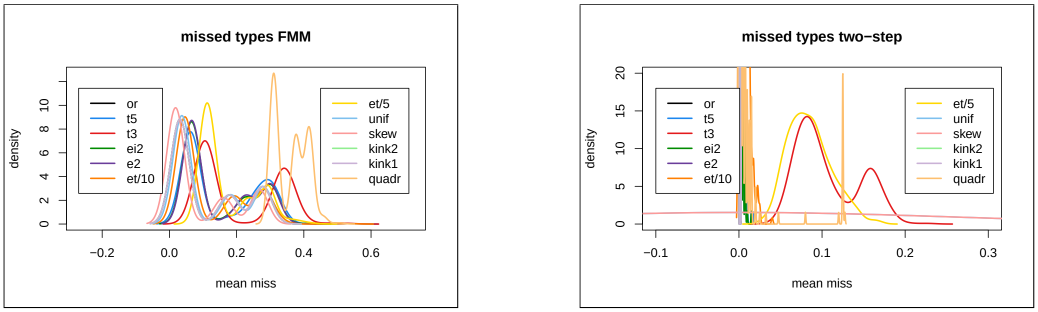

Figure 6: Fraction of Misclassified Types
left panel: finite mixture; right panel: two-steps

A final performance criterion is the bias in estimating the treatment effect, conditional on type. In the finite mixture model, the predicted treatment effect can be directly read off the coefficient for $\theta \cdot t$ (see Table 2). In the two-step estimator, for the first

data generating processes, and hence the density extremely large at 0. The flat line for the skewed data generating process results from the fact that misclassification is almost perfectly normally distributed around mean 0 in this case, with minimum and maximum also very close to 0.

type it is given by $t \cdot \hat{k}_2$. For the remaining types it can be calculated as $t \cdot \hat{k}_4 - t \cdot \hat{k}_3$, $t \cdot \hat{k}_6 - t \cdot \hat{k}_5$ and $t \cdot \hat{k}_8 - t \cdot \hat{k}_7$, respectively. This prediction can be compared with the population effect defined by (4).

As Figure 13 in the Appendix shows, in this respect the finite mixture model and the two-step estimator each have comparative advantages and disadvantages. Depending on type and on the specification of the data generating process, the distributions of biases observed in the 500 samples per data generating process differ substantially. The mean of these biases per condition is not very informative: underestimations of the treatment effect and overestimations may cancel out. The range can also be misleading if there are a few outliers in the respective simulation.[15] This is why, in Table 5, I report the width of the interquartile range, i.e. the difference between the bias at 75% of the respective distribution and the bias at 25%.

The two-step estimator takes type information from the development of individual choices over time. The shorter the panel, the less information the estimator has. As Table 5 shows, a panel of length five is still reasonably long. For all types, the two-step estimator outperforms the finite mixture model, and the misestimation of the treatment effect still remains rather small. Yet misestimation jumps up if the panel has only length three, and for three of the four types, the two-step estimator performs more poorly than the finite mixture model. Results look similar if the development of choices over time is only observed with error, i.e. if there is randomness in the slopes. Finally the two-step estimator gets the treatment effect for one of the types wrong if a quadratic term in the development over time is neglected, but in this condition the finite mixture model gets the treatment effect for two other types wrong.

However the two-step estimator clearly dominates the finite mixture model in the remaining conditions, i.e. in the original data generating process, with a panel of length five, with more pronounced noise at the individual level or in the residual error. This also holds if error is taken from a uniform or beta distribution, or if there is a kink in the development of choices of one type over time. This is interesting as prediction accuracy and the identification of types had still been good in these conditions when using a finite mixture model.

The two-step estimator is not a panacea. In some conditions, the finite mixture model draws a more accurate picture of the population effect. Yet the simulations suggest that the two-step estimator is at least as good as the finite mixture model if the information about the development of individual choices over time is sufficiently rich, and sufficiently accurate. It will be for the experimenter to decide whether she is confident that these conditions are fulfilled in her data. Actually if she suspects heterogeneity in the treatment effect, she can adjust the design and have participants decide for more periods.

[15] In Figure 13 the x-axis of several plots is cut at the extremes for this reason.

		type1		type2		type3		type4	
		FMM	two	FMM	two	FMM	two	FMM	two
1	original	.056	.03	.137	.030	1.994	.029	.963	.030
2	panel length	.215	.086	.574	.095	2.089	.096	.926	.091
3		.604	1.853	1.528	10.498	2.315	2.176	1.032	2.291
4	noise	.115	.031	.278	.038	2.036	.030	.937	.030
5		.131	.006	.305	.059	2.053	.057	.957	.059
6	random coefficient	.138	5.976	.106	2.069	2.080	.152	.905	2.031
7		.222	.668	.100	10.469	1.924	3.623	1.151	2.365
8	distribution	.058	.029	.180	.029	2.004	.029	.975	.029
9		.040	.029	.073	.030	1.995	.032	.969	.026
10	kinked	.056	.030	.137	.030	1.994	.029	.963	.030
11		.056	.030	.137	.030	1.994	.029	.963	.030
12	dynamics	.087	2.196	.064	.044	1.410	.033	1.613	.032

Table 5: Misestimation of Treatment Effect
Width of Interquartile Range of Distribution of Bias Estimates

6 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 (7).[16]

$$\pi_i = 20 - c_i + .4 \sum_{j=1}^{4} c_j \tag{7}$$

In (7) π_i is payoff, c_i is the contribution a participant makes to the public good of a group of size $J = 4$. As $.4 < 1$, it is individually rational to keep the endowment. Yet as $4 \cdot .4 = 1.6 > 1$ it is socially rational that all group members contribute their entire endowments. The novelty is the use of 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. 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

[16] As I have used k throughout the paper to characterize types, with a slight abuse of notation in this equation I use j for **any** member of the group, including i.

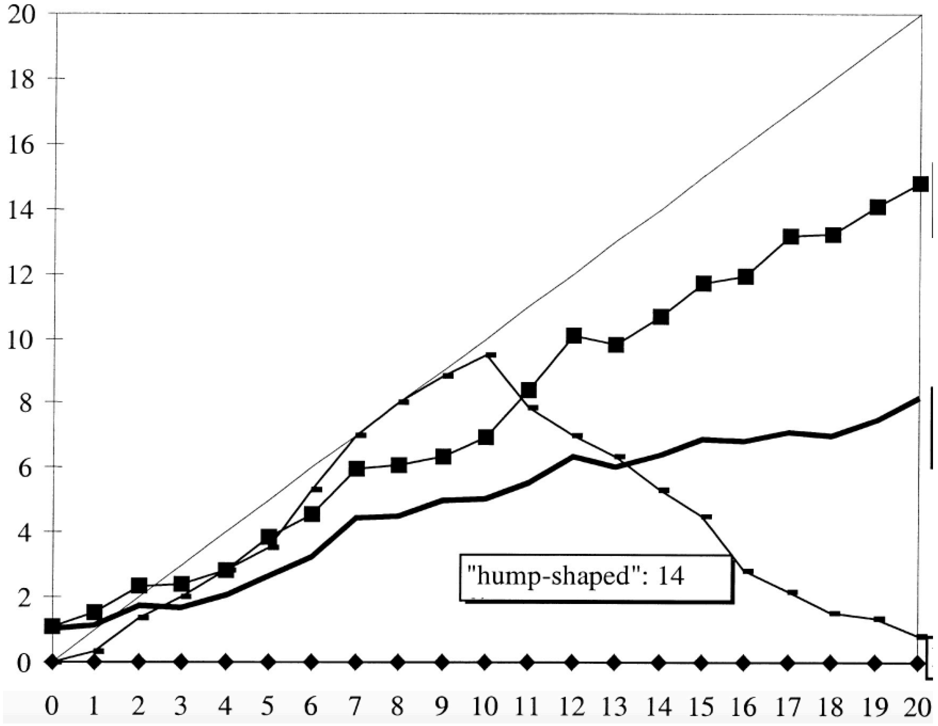

Figure 7: Cooperation Types, Fischbacher Gächter Economics Letter 2001, Fig.1
x-axis c_{-i}, y-axis c_i solid line is overall mean

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 7.

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 (8)

$$c_{il} = \beta_0 + \beta_1 c_{-i} + u_i + \epsilon_{il} \tag{8}$$

The strategy method exposes participants to a within-subjects design. Treatment consists of the number of tokens the remaining group members on average contribute to the joint project. There are $L = 21$ possibilities, ranging $l = c_{-i} \in \{0, ..., 20\}$. As the participant in question stays the same, a specification is in order that filters out unobserved individual idiosyncrasies with the random effect u_i. If one estimates (8),

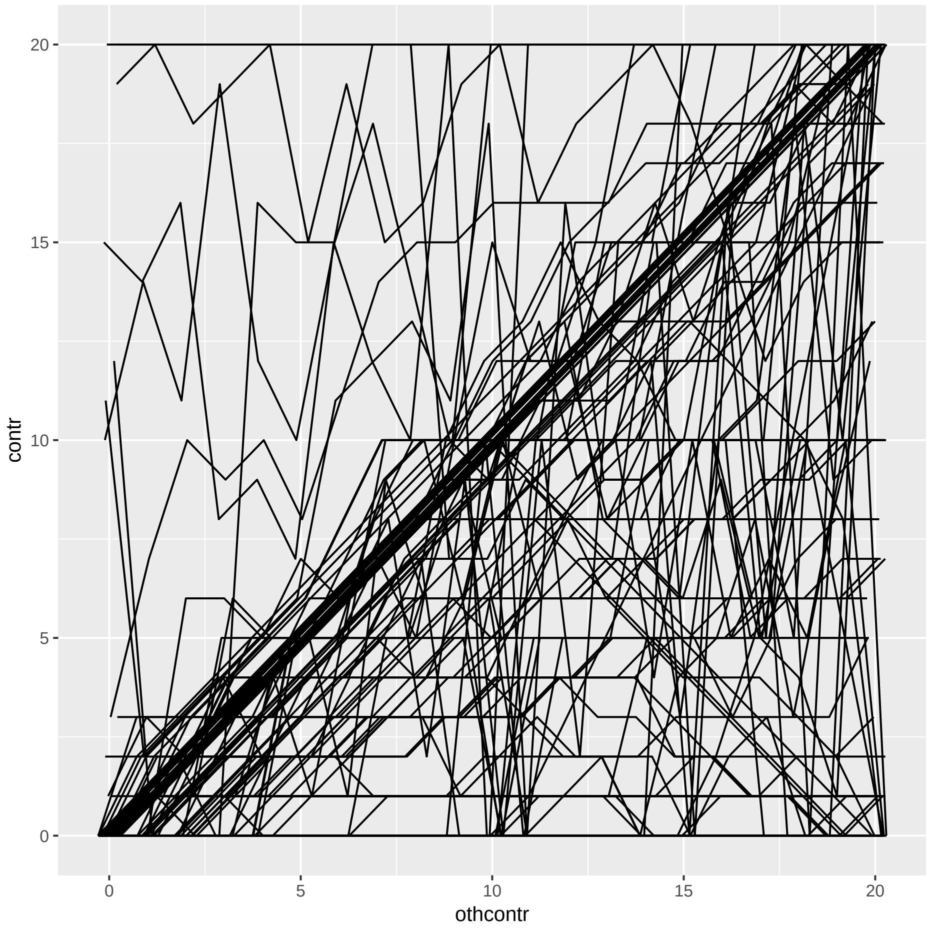

Figure 8: Fischbacher & Gächter Distribution of Raw Data.
Each participant represented by a separate line
Thickness of lines represents frequency

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 8 shows clear heterogeneity. Inspecting the figure suggests one relatively clear type: Perfect conditional cooperators. Yet the choices of many participants look very different. Attempts at estimating a finite mixture model fail as the model does not converge, 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.

Figure 9 collects the results. It represents mean choices per type. The upper left panel is resulting from, separately for each participant, regressing c_{il} on c_{-i}. As Figures 7 and 8 suggest the possibility of a non-linear relationship, the upper right panel of Figure

	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 6: Type Space in Fischbacher Gächter

7 is derived from local (per participant) regressions of c_{il} on $c_{-i} + c_{-i}^2$. The former exercise yields 7 distinct types, the latter 8. As Table 6 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 in Figure 11 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 **??**). 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 9. For each type, it aggregates over conditional choices, separately for each possible (mean) unconditional choice. Whether or not local regressions include a quadratic term (upper right and upper left panels), the following three types are evident: 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 unconditional choices; a type that matches very low unconditional choices, but then levels off; a type that is selfish if unconditional choices are

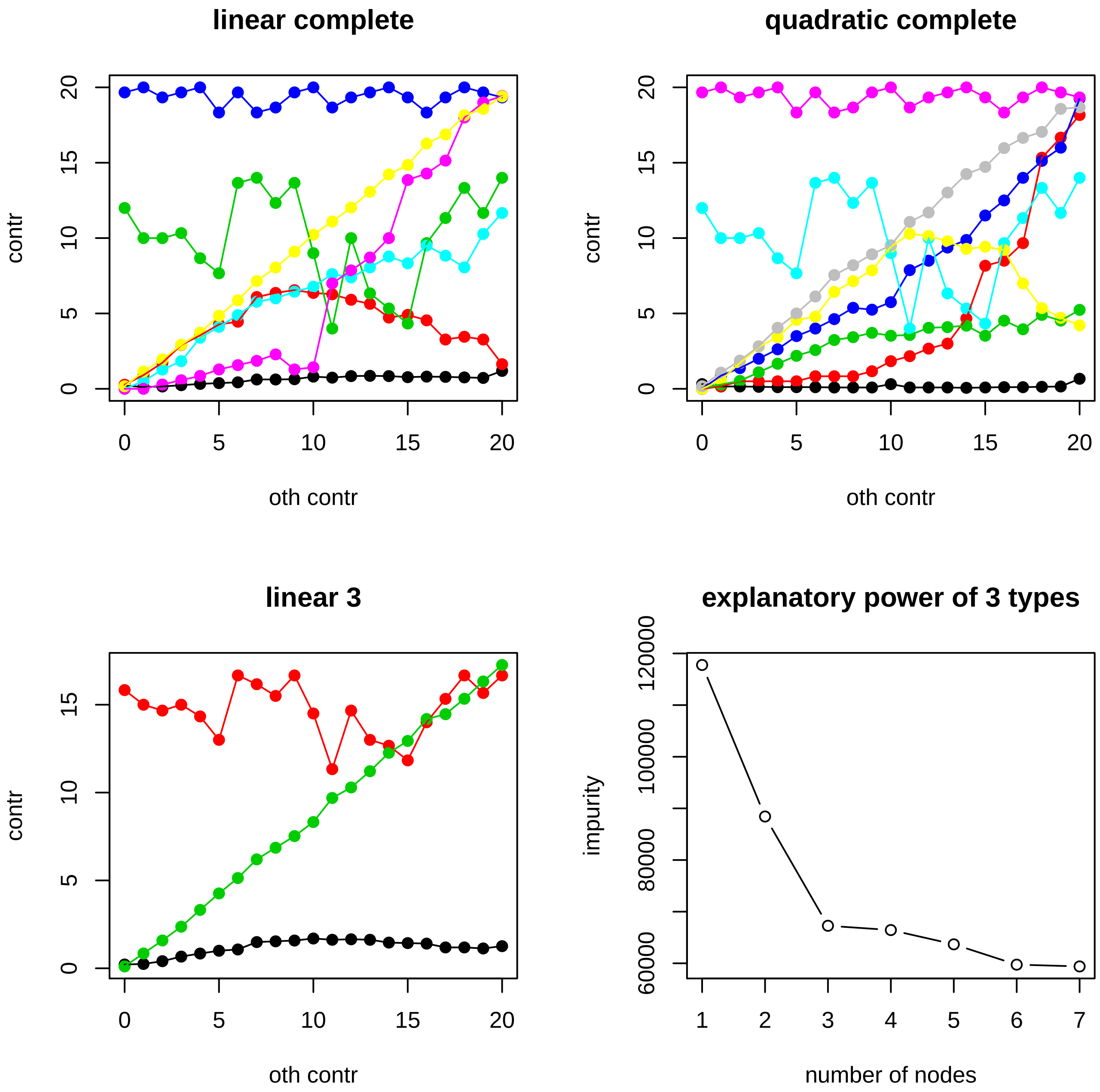

Figure 9: Types Induced by CART

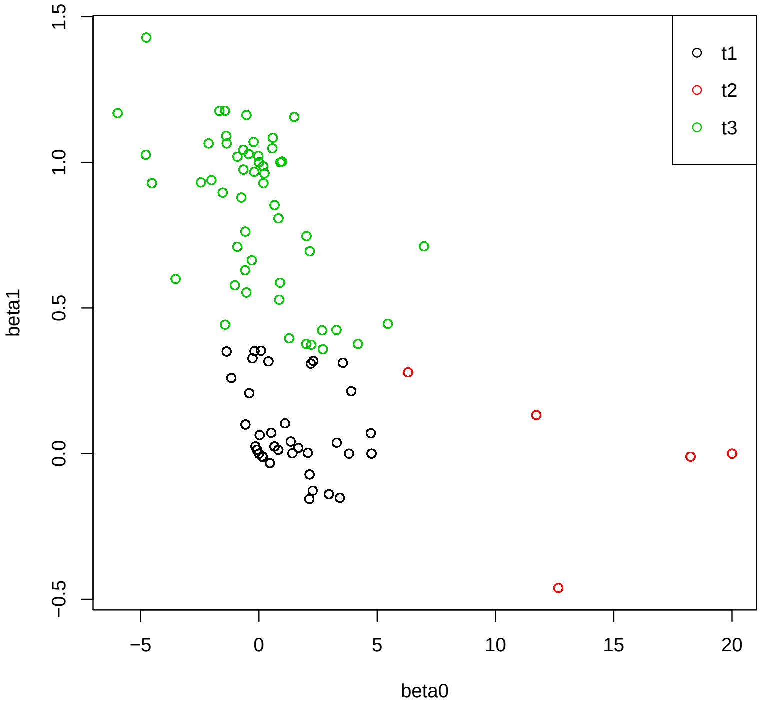

Figure 10: Coefficients of Local Regressions for Tree with Three Types

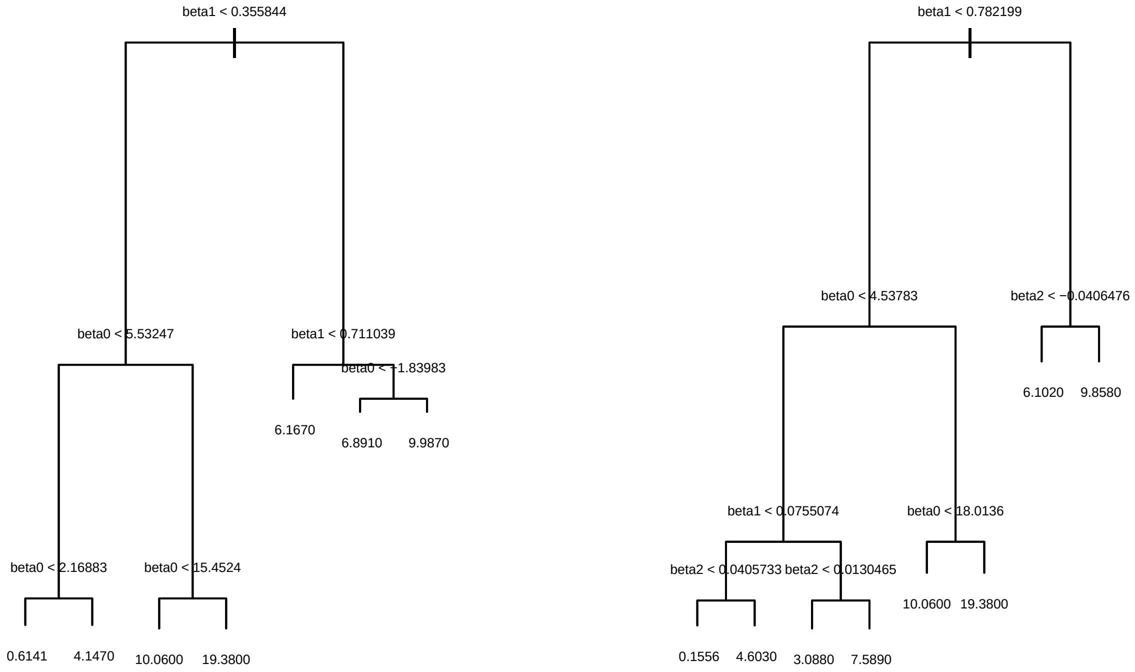

Figure 11: Trees Induced by Local Regression with Only Linear / also Quadratic Term

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 9 shows that the loss in precision is low if one only allows for three types.[17] 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. Figure 10 shows in which ways CART defines these three types, simultaneously using the constants (β_0) and the slopes (β_1) from the local regressions.

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.

7 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) one must make assumptions about a two-dimensional distribution; (c) 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; (d) finite mixture models typically use two-dimensional maximum likelihood estimation; the datasets from experiments may be 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

[17] Impurity is a measure for imprecision, number of nodes refers to the nodes in the (pruned) regression tree.

each participant, I estimate a local (per participant) 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). CART is a non-parametric procedure, and hence does not require distributional assumptions (b). The final step of the procedure can easily handle random and mixed effects models, as at this point type need no longer be estimated (c). 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 (d).

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 (clusters) 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 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 (for background on bagging and boosting see James et al., 2013).

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. One approach is a finite mixture model. One simultaneously estimates the type space, and reactions to treatment conditional on type. This paper proposes an alternative simple and robust method, provided the experiment is repeated. Simulations show that it is more accurate than the finite mixture model provided type only varies between subjects, and information about the development of choices over time is sufficiently rich and precise. Experimenters have an alternative to the finite mixture model at their disposition.

References

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn. A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94(2):481–498, 2012.

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.

Mahmoud A El-Gamal and David M Grether. Are people bayesian? uncovering behavioral strategies. *Journal of the American statistical Association*, 90(432):1137–1145, 1995.

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.

William H Greene. *Econometric analysis.* Pearson Education India, 2003.

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.

Kevin M Murphy and Robert H Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379, 1985.

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.

Terry M Therneau, Beth Atkinson, Brian Ripley, et al. rpart: Recursive partitioning. *R package version*, 3(3.8), 2010.

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.

Daniell Toth and John L Eltinge. Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496):1626–1636, 2011.

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 1: Code in R

step 1: define data

Dataset dat is assumed to be a panel with crosssection uid $\in \{1, ..., N\}$. Each individual is observed in periods per $\in \{1, ..., T\}$. Dependent variable y is observed, as well as treatment treat $\in \{0, 1\}$.

step 2: initialize vector of coefficients

```
beta0 <- rep(0,N)
beta1 <- rep(0,N)
uid <- 1:N
rdat <- data.frame(uid = uid, beta0 = beta0, beta1 = beta1)
```

step 3 - 4: write and execute local regression, collect output in separate dataset

```
loc <- function(u) {
lmu <- lm(y[uid == u] ~ per[uid == u], data = dat)
rdat$beta0[uid == u] «- lmu$coefficients[1]
rdat$beta1[uid == u] «- lmu$coefficients[2]
}
invisible(lapply(1:N,loc))
```

step 5: merge with main dataset and keep variables needed for tree construction

```
mdat <- merge(rdat, dat, by = "uid")
trdat <- mdat[,c(1:4)]
```

step 6 - 7: fit classification tree and define optimal depth

```
library(tree)
tytree <- tree(y ~ beta0 + beta1, data = trdat)
plot(tytree)
text(tytree)
```

step 8 - 9: assign type to participant, and split types into treated and untreated cases

note: correct code depends on structure of tree; following is a simplified example

```
mdat$ttype <- ifelse(mdat$beta0 < -3 & mdat$beta1 < 2, 1,
ifelse(mdat$beta0 < -3 & mdat$beta1 >= 2), 2, 3)
mdat$trtype <- ifelse(mdat$ttype == 1 & mdat$treat == 0, 1,
ifelse(mdat$ttype == 1 & mdat$treat == 1, 2,
ifelse(mdat$ttype == 2 & mdat$treat == 0, 3,
ifelse(mdat$ttype == 1 & mdat$treat == 1, 4,
ifelse(mdat$ttype == 3 & mdat$treat == 0, 5, 6 )))))
```

step 10: estimation conditional on type (and treatment)

```
library(plm)
mlretype <- plm(y ~ as.factor(trtype)*as.numeric(per), data = mdat,
index = c("uid", "per"), model = "random")
summary(mlretype)
```

Appendix 2: 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 (1)

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 12.

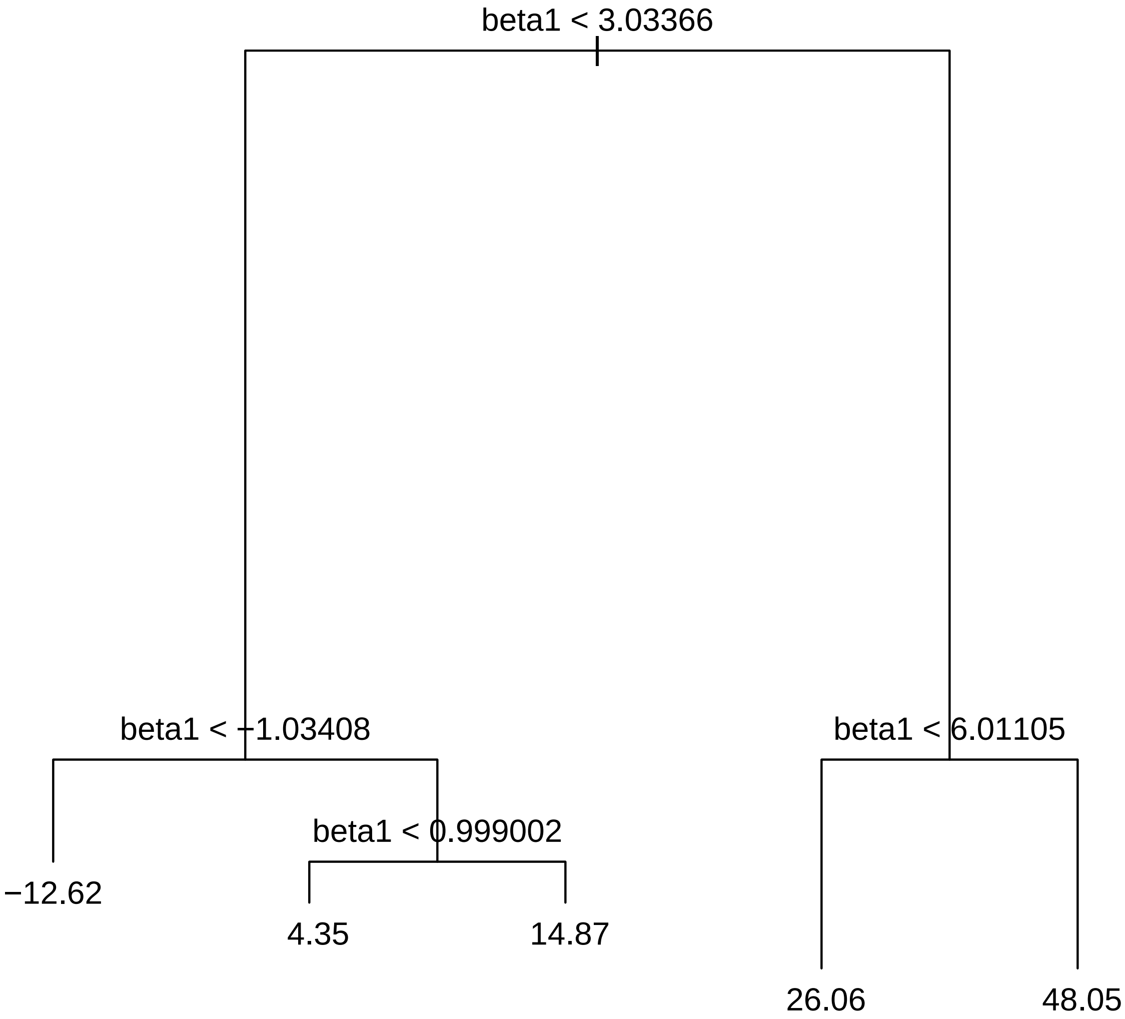

Figure 12: Tree Induced by Precision Weighted Local Regressions

Appendix 3: Graphical Representation of Misestimation of Treatment Effect Conditional on Type

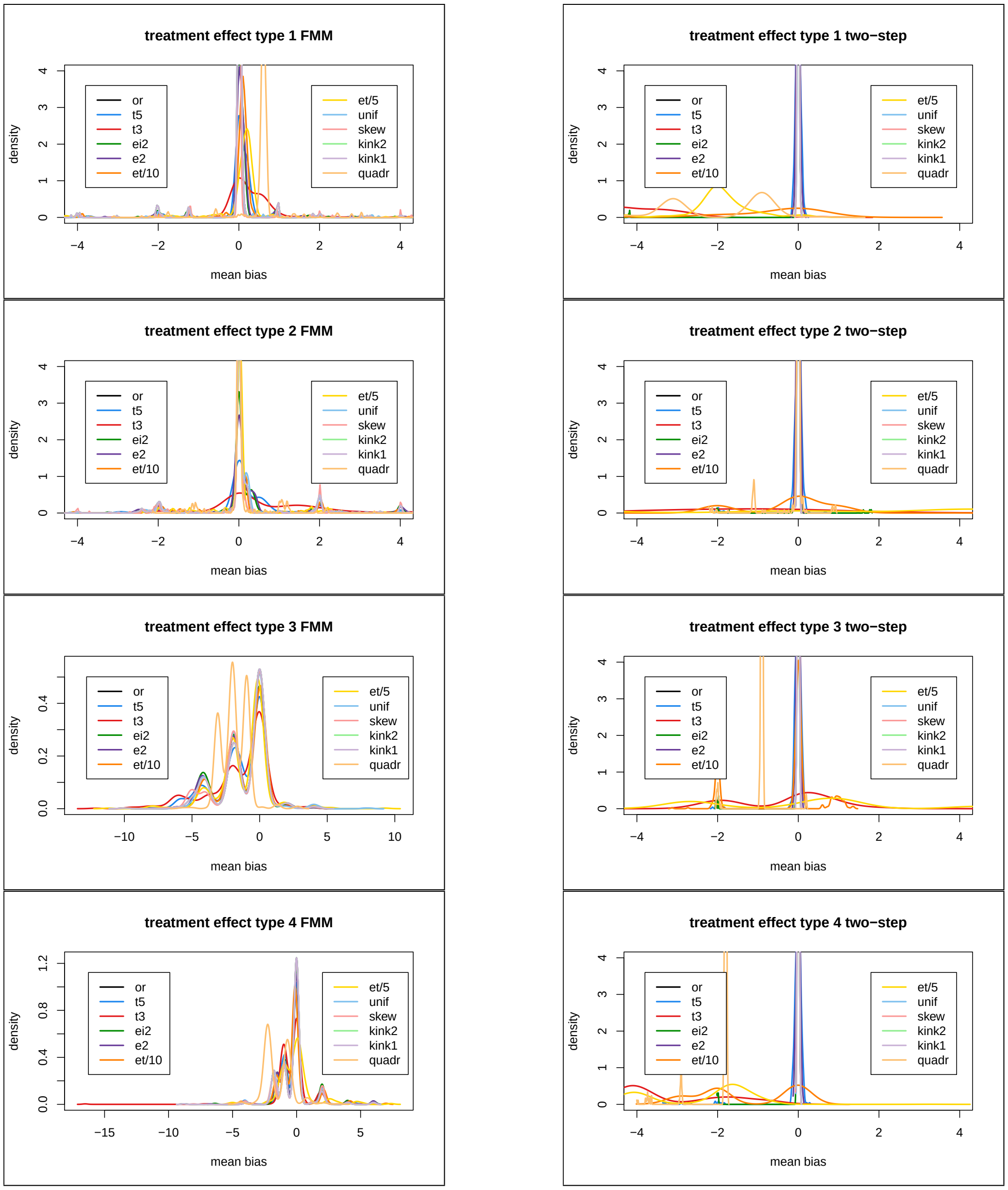

Figure 13: Bias in Estimation of Treatment Effect Conditional on Type
left panel: finite mixture; right panel: two-steps