Bayesian Statistics Project

Nonparametric Bayesian Analysis of Simulated Heterogenous Treatment Effect

César Roaldès & Morgane Hoffmann

January 2019

1 Introduction

In this project we will use as a reference paper "*A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation*" from Taddy et al. (2016) [3]. Our main goal is to apply a *distribution free Bayesian nonparametrics analysis* presented in this paper on simulated data in order to estimate average and heterogeneous treatment effects for randomized control trials. These estimations will be carried out using two different statistical approaches, the OLS regression and decision trees, the results of which will then be compared. As this paper is based on real data we propose in this project a simulation study in order to check the main results.

2 Data Generating Process

Following Taddy et al (2016), we represent the Data Generating Process (DGP) through a probability mass function over a large but finite number of possible data points $\mathbf{z}$ (the combination of responses, covariates and treatments).

$$g(\mathbf{z};\boldsymbol{\theta}) = \frac{1}{|\boldsymbol{\theta}|}\sum_{l=1}^{L}\theta_l \mathbf{1}_{\{\mathbf{z}=\zeta_l\}} \tag{1}$$

- $\mathcal{Z} = \{\zeta_1, ..., \zeta_L\}$ represents the support of the DGP.
- $\boldsymbol{\theta}$ are random weights $\theta_l \geq 0$.
- $|\boldsymbol{\theta}| = \sum_l |\theta_l|$

The conjugate prior for the normalized vector $\boldsymbol{\theta}/|\boldsymbol{\theta}|$ is a Dirichlet distribution with a single parameter of concentration α. We note $\mathbf{Z} = [z_1, \ldots, z_n]'$ the observed sample, hence, the posterior distribution for $\boldsymbol{\theta}/|\boldsymbol{\theta}|$ is :

$$\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|}|\mathbf{Z} \propto \prod_{i=1}^{n}(\theta_i/|\boldsymbol{\theta}|)^{\alpha} \prod_{i=n+1}^{L}(\theta_i/|\boldsymbol{\theta}|)^{\alpha-1} \tag{2}$$

Using the non-informative limit prior that arises as $\alpha \to 0$ yields to the massive computational convenient characteristic of the Dirichlet distribution : the posterior distribution on $\boldsymbol{\theta}/|\boldsymbol{\theta}|$ is equivalent to independent exponential posterior distribution on each unnormalized vector component $\theta_l|\mathbf{Z} \overset{\text{ind}}{\sim} \text{Exp}(\alpha + \mathbf{1}_{\{l\leq n\}})$. In practice, following Rubin (1981) [4], we obtain a sample of the posterior on the statistic of interest $S(\boldsymbol{\theta})$ (for instance a posterior DGP mean or a CART predictor) through Bayesian Bootstrapping (BB). This method is presented in the Algorithm 1.

Algorithm 1 Bayesian Bootstrap

for b = 1, ..., B **do**
  draw $\theta_i^b \overset{\text{iid}}{\sim} \text{Exp}(1), \quad i = 1, \ldots, n$
  compute $S_b = S(\boldsymbol{\theta}_b)$
end for

3 Interest of the Nonparametric Bayesian Approach in A/B Testing Context

For a long time, available datasets were too small to explore heterogeneity of treatment effects, but recently, with the increase in the amount of data generated, this type of experimental settings have been made possible. Frequentist analysis usually focus on testing and detecting the existence of heterogeneous treatment effects. Bayesian methods can be a good alternative to the frequentist approach in this case because it naturally embed heterogeneity in the posterior distribution. The goal is then to focus on characterizing this heterogeneity through posterior uncertainty. The second main difference of Bayesian method is the inclusion of a prior information to the model. In our case however, the authors do not make advantage of this characteristic and use the prior distribution more as a mathematical tool than as a philosophical concept.

A/B experiments are large-scaled randomized experiments commonly made by firms in order to study the impact of a new advertising or marketing policy on consumer purchases. The data collected during A/B experiment show unusual features that require to be taken into account. First, sample sizes are bigger. Second, effect's sizes are small and thus response's standard deviations are relatively big. Finally, the distribution of the outcome variable of interest tends to be complex to model with parametric distribution. The "*bad behavior*" exhibited by those distributions can be summarized by the following four usual features :

- a majority at zero (most users do not make any transactions during the experiment);
- a long and fat right tail (heavy spend by a minority of users);
- density spiked at psychological thresholds;
- a correlated variance with both treatment and sources of treatment heterogeneity.

Using a distribution-free Bayesian nonparametrics analy-

sis allows to make minimal assumptions on the DGP, aside from independence across observations. This flexibility can be useful in case of doubts on the linearity of the treatment effect. Moreover, this method is easy to implement and the posterior convenient to sample. It is particularly true in the big data context where MCMC techniques can be costly or even impossible to apply. Moreover, using the BB provides an interpretation of the resulting B statistics $S(\boldsymbol{\theta}_b)$ as the posterior distribution of the statistic S instead of the estimated sampling distribution provided by the nonparametric frequentist bootstrap.

4 Data Simulation

Our principal objective is to simulate a dataset exhibiting similar characteristics as one could observe while proceeding an A/B testing. Due to our computational resources limitation, we are constrained to restrict the size N of our dataset to one million observations and use a set of 10 simulated random features. This set contains 2 binary variables and 8 continuous variables drawn from the probability distributions detailed below. Then, we randomly assign half of our N individuals to the *treatment group* ($T_i = 1$) and the other half to the *control group* ($T_i = 0$), the treatment is therefore completely independent of the outcome variable y. We construct a treatment function $\tau(X)$ which depends on two variables X_{sex} and $X_{\text{spend}}(\in X_{\text{continuous}})$. For sake of clarity, we interpret those variables as the gender and the amount spent during the last month. The other features are only used to create the DGP but will have no impact on the treatment effect. We denote X_{new} a binary variable (interpreted as the indicator for a new consumer) not linked to the treatment effect but will however be used in order to check how our estimators will exploit this variable to estimate its effect. In order to reproduce the fact that a minority of user spend massive amounts we choose a Pareto distributed error for the treatment function. Finally, one of the problems arising when analysing A/B testing data is the large amount of individuals who do not purchase during the experiment. In order to replicate this feature we assigned randomly $y_i = 0$ to 70% of our observations and end up with a mixed model for y_i.

Finally we construct the outcome variable as follow:

$$
\begin{aligned}
&X_{\text{sex}} \sim \mathcal{B}(0.7),\ X_{\text{new}} \sim \mathcal{B}(0.8) \\
&X_{\text{continuous}} \sim Exp(1) \\
&T_i \sim \mathcal{B}(0.5)
\end{aligned}
$$

$$
\begin{aligned}
&\tau(X) = \mu + \gamma_1 X_{\text{spend}} + \gamma_2 X_{\text{sex}} + \nu, && \nu \sim Par(0.5, 2) \\
&\nu(c) = \alpha + X\beta + e, && e \sim \mathcal{N}(0, 1) \\
&\nu(t) = \alpha + X\beta + \tau(X) + e \\
&y_i = 0.7\delta + 0.3\left[T\nu(t) + (1-T)\nu(c)\right]
\end{aligned}
$$

With β a vector of length 10 equal to $[2, \ldots, 2]'$, $\alpha = 1$, $(\mu, \gamma_1, \gamma_2) = (1, 1, 1)$, and δ the Dirac delta function.

5 Results

Taddy et al. (2016) presents two different approaches to estimate treatment effects, one using *weighted population OLS regressions* and the other based on *bayesian regression trees*. The analysis of the Average Treatment Effect (ATE) by both methods is developed in 5.1 before focusing on the Heterogenous Treatment Effects (HTE) in 5.2.

5.1 Average Treatment Effect

The ATE measures the difference in mean between a treated and control population in randomized experiments.

5.1.1 ATE by Weighted OLS Regression

In this method, our continuous covariates are transformed such as each x_{ij} element indicates whenever that continuous variable is greater than or equal to the 20th, 40th, 60th and 80th percentile. Binary covariates are not pre-processed. Moreover we set $x_{i1} = 1, \forall i$ for the intercept.

Using our nonparametric Bayesian perspective we define the group population OLS projections which is a weighted OLS with weights drawn as presented in the algorithm 1:

$$\beta_d = (X_d' \Omega_d X_d)^{-1} X_d' \Omega_d y_d \tag{3}$$

Where $\Omega_d = Diag(\boldsymbol{\theta}_d)$ and $\boldsymbol{\theta}_d$ is the sub-vector of weights for individuals in group d ($d = c$ for the control group and t for the treatment group). This statistic is similar to the frequentist White correction matrix. From this statistics we can derive the posterior of the difference $\beta_t - \beta_c$.

Following the method in section 2 we compute 3 different statistics for the ATE :

- The usual measure of ATE for a randomized control trial that is, the difference in group means : $\bar{y}_t - \bar{y}_c$
- A modified metric that conditions upon the informative in covariates $\mathbf{x}$, the *regression-adjusted ATE* : $\bar{\mathbf{x}}'(b_t - b_c)$
- A bayesian nonparametric analogous of the previous statistic : $\mu_x'(\beta_t - \beta_c)$ where $\mu_x = X'\boldsymbol{\theta}/|\boldsymbol{\theta}|$

5.1.2 ATE by Bayesian Regression Trees

Using regression tree, and more specifically CART algorithm [2], allow us to make predictions that are nonlinear in the original covariates and does not imply homoscedasticity's assumption. Taddy et al. introduced the methodology behind the Bayesian Forests in a previous article *"Bayesian and Empirical Bayesian Forests"* (2015) [5]. The latter suggests to fit weak learners CART trees over realizations of the DGP model (1), meaning that for each split, the impurity minimization is weighted by $\boldsymbol{\theta}$ over our sample. Like the usual CART algorithm, for a node η, a binary split is recursively carried out according to the feature j at the location x_j^* such as the two resulting child nodes are $left(\eta, j, x_j^*) = \{i : x_{ij} \leqslant x_j^*, i \in \eta\}$ and $right(\eta, j, x_j^*) = \{i : x_{ij} > x_j^*, i \in \eta\}$ until a stopping rule is encountered. The variation induced by the bayesian CART algorithm is that given a random DGP realization

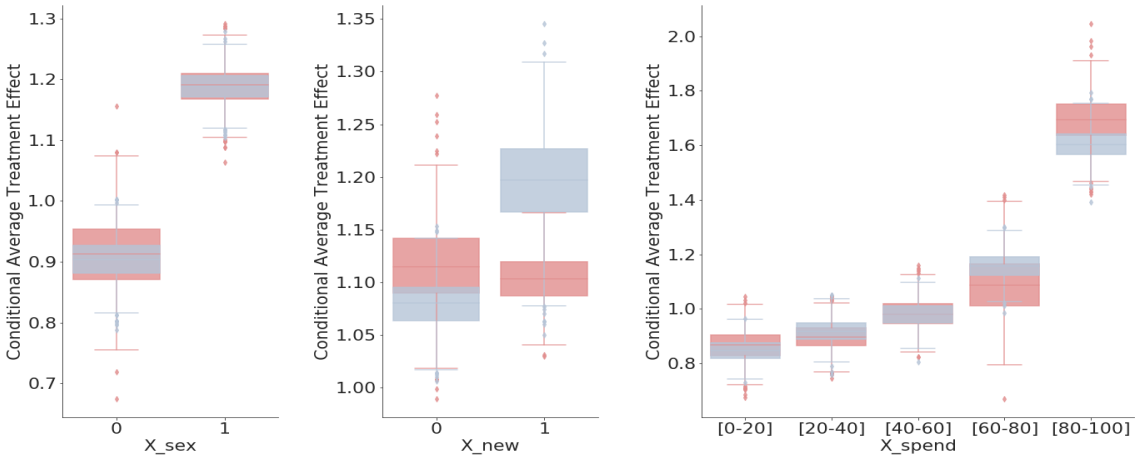

Figure 1: ATE conditional on variable j being in set $\mathcal{X}$

$\boldsymbol{\theta}$, the choice of the feature x_j and the splitting point x_j^* is driven by the minimization of

$$\varepsilon_{left(\eta,j,x_j^*)}(\boldsymbol{\theta}) + \varepsilon_{right(\eta,j,x_j^*)}(\boldsymbol{\theta}),$$

where for a generic node $\mathcal{S} \subseteq \{1,\dots,n\}$, the impurity is

$$\varepsilon_{\mathcal{S}} = \sum_{i\in\mathcal{S}} \theta_i (y_i - \mu_{\mathcal{S}})^2 \quad \text{with} \quad \mu_{\mathcal{S}} = \boldsymbol{y}'_{\mathcal{S}}\boldsymbol{\theta}_{\mathcal{S}}/|\boldsymbol{\theta}_{\mathcal{S}}|.$$

The posterior over trees is sampled via the Bayesian Bootstrap outlined in Algorithm 1, labeled by Taddy et al. as *Bayesian Forest* (BF). We manually implement the BF within the `scikit-learn` library on Python by modifying the `forest.py` script. Drawn on the frequentist nonparametric bootstrap used by the Random Forests, we switch the discrete random weights vector drawn from a Multinomial distribution with probability of $1/N$ of length N by independent Exp(1), reproducing our DGP of section 2. It is worth to notice that unlike Random Forest, CART algorithm is applied without any random variable sub-setting, so the variability in the resulting predictions is due to posterior uncertainty about the DGP.

Just as in the weighted OLS regresion, we build *population BF*, and construct a BF for each population in our sample. Once both fitting process are performed, each DGP realization provides an ATE's CART prediction defined by :

$$\hat{y}_t - \hat{y}_c = \frac{1}{|\boldsymbol{\theta}|}\sum_{i=1}^{N} \theta_i(\hat{y}_t(\mathbf{x}_i - \hat{y}_c(\mathbf{x}_i)),$$

where $\hat{y}_d$ is the prediction of a tree in the BF of population d. Notice that the random $\boldsymbol{\theta}$ plays a role in both the fitting process and the ATE's prediction. We then compute the posterior mean[1] and standard deviation of $\hat{y}_t - \hat{y}_c$ in our BFs.

5.1.3 Comparison of ATE's predictions

Like Taddy et al.(2016), the results obtained through the different approaches are really close in both means and variances. This is mainly due to the way we simulated the data where effects are small as highlighted by the paper. Bayesian trees method does not seem to perform better in terms of standard deviation.

Estimator	$\bar{y}_t - \bar{y}_c$	$\bar{x}'(b_t - b_c)$	$\mu'_x(\beta_t - \beta_c)$	$\hat{y}_t - \hat{y}_c$
Mean	1.1039	1.1041	1.1099	1.1056
SD	0.0212	0.0209	0.0198	0.0215

Table 1: Posterior means and standard deviations for ATE statistics according to different measures

5.2 Heterogenous Treatment Effect

The HTE analyzes the difference of treatment impact when we consider different strata of our sample. This difference is directly obtained in the weighted OLS regression by the coefficient's obtained over our two population. Hence, this section focus on HTE estimation using BF methods.

5.2.1 HTE by Bayesian Forest

The way we defined BF allows us to consider the latter as a posterior over CART predictors, and thus, the average leaf value associated with a given $\mathbf{x}$ as the posterior mean prediction rule. For each CART in the BF, the leaf node $\mathcal{S}$ containing $\mathbf{x}$ associates $\hat{y}_d(\mathbf{x}) = \mu_{\mathcal{S}} = \mathbf{y}'_{\mathcal{S}}\boldsymbol{\theta}_{\mathcal{S}}/|\boldsymbol{\theta}_{\mathcal{S}}|$. As a result, we are able to predict the treatment effect of each individual using a CART fitted within our population BF as $\hat{y}_t(\mathbf{x}_i) - \hat{y}_c(\mathbf{x}_i), \forall i$. From this statement, we are able to get the ATE conditional on variable j being in set $\mathcal{X}$ as :

$$\hat{y}_t^j(\mathcal{X}) - \hat{y}_c^j(\mathcal{X}) = \frac{\sum_{i:x_{in}\in\mathcal{X}} \theta_i(\hat{y}_t(\mathbf{x}_i) - \hat{y}_c(\mathbf{x}_i))}{\sum_{i:x_{in}\in\mathcal{X}} \theta_i}.$$

This equation gives us the treatment effect associated across the modalities[2] taken by the variable under study, and thus, an estimation of the HTE. Figure 1 shows the results obtained using the weighted OLS regression (in light blue) as described in the previous section and the population BF approach (in light red). Both HTE estimations have been realised on one million observations. Among those three variables, only X_{sex} and X_{spend} affect the response y of an individual, unlike X_{new} that has no additional effect if treated. We find the expected results

[1] This posterior mean deffer from the difference between the whole BF by the fact that each difference on prediction's CART is weighted by a new $\boldsymbol{\theta}$

[2] For discrete variables. We use quantiles for continuous ones.

arising from our simulation for each variable. The mean spread between the two modalities of X_{sex} is around 0.3 and correspond to the treatment's coefficient associated to the variable, while X_{spend} display the exponential shape of the Exp(1) random variable.

5.2.2 TOT method

Athey and Imbens (2015) [1] introduced the *Transformed Outcome Tree (TOT) method* which is CART prediction over a transformed response y^* in order to have the transformed response expectation equals to the treatment effect. This methodology derived from the Neyman-Rubin causal model, which considers that each observation i has two potential outcomes, $v_i(\text{c})$, observed if the individual is in control group ($T_i = 0$), and $v_i(\text{t})$ if he is treated ($T_i = 1$). By defining y_i^* as

$$y_i^* = y_i \frac{T_i - q}{q(1-q)} \quad \text{with} \quad q = \mathbb{P}(T_i = 1),$$

the expectation is equal to

$$\begin{aligned} \mathbb{E}[y_i^* | v_i] &= v_i(\text{t}) q \tfrac{1-q}{q(1-q)} - v_i(\text{c})(1-q)\tfrac{q}{q(1-q)} \\ &= v_i(\text{t}) - v_i(\text{c}). \end{aligned}$$

It is worth underlining the fact that TOT's leafs need to have a large number of observations for the empirical expectation to converge. Trees partition observations in regions of response homogeneity and associate with each partition the average of the responses within it. This allows us to directly predict the treatment effect since trees are fitted on the response y_i^*.
This last point is crucial. Indeed, choosing the splitting point x_j^* minimizing the variance within the two resulting subgroups allows us to interpret the variable x_j as the variable that most impacts the treatment effect. The structure's analysis of each tree among the BF gives us the posterior probability of decision nodes for a given tree.

Figure 2 shows the trunk of a tree with the posterior probability that the variable used in each decision nodes emerges at or above its current depth whereas Table 2 regroups those probabilities. This tree's root node splits our dataset using X_{spend}, spotting that this variable has the strongest impact on the treatment effect, conditionally upon being at or below 1.156. This result is validated by the fact that the second variable linked with the treatment effect X_{sex} is a binary variable. Because both variables have the same coefficient, splitting X_{spend} above 1 leads to a greater impact on the impurity minimization[3]. This variable is systematically selected as the root node, which attests to its certainty. On the left side of the tree, the next split is realized upon the variable X_{sex}, as theoretically expected because the splitting variable is chosen on a sub-sample of individuals having small values for X_{spend}, and thus a small treatment effect associated with the latter. However, the emergence of X_{sex} at this depth or above is more uncertain and happens only in 42% of trees. On the other side, a second consecutive split is performed on the same variable X_{spend}. Notice that the difference between the splitting value of the variable is again greater than 1, exceeding the maximum value one could get with

[3]And for completeness, the large amount of individuals having $X_{\text{spend}} > 1.156$

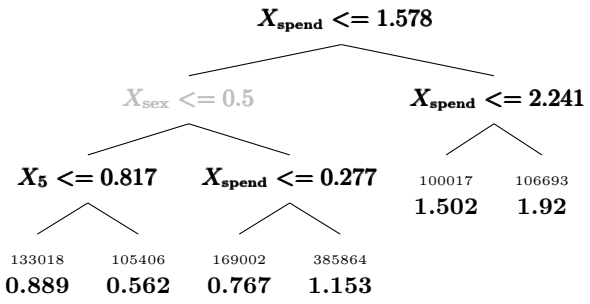

Figure 2: Sample TOT tree trunk. The right-hand children contains data for which the split condition is true. Decision nodes are coloured for the posterior probability : light gray for $p < \frac{1}{2}$ and **black** for $p > \frac{1}{2}$. The sample sizes and the value of y^* are indicated in the terminal leaf.

	depth in tree		
	1	≤ 2	≤ 3
X_{spend}	1	1	1
X_{sex}	.00	.42	.68
X_5	.00	.07	.47

Table 2: Posterior probabilities of the TOT algorithm splitting at or above 1 to 3 depth on each of the variables in the corresponding sample TOT tree. We presented only values for the variable presented in the tree above

X_{sex}. Surprisingly though, the next value used to split the resulting left child is X_4 despite the fact that the value is not linked with the treatment, but with low probability.

6 Issues Encountered And Further Developments

During this project we faced several issues. One of the main difficulty encountered was to replicate A/B experiment data characteristics while being able to verify that the results were correct and ensure that the methodology was well implemented. Moreover, we had to deal with the computational resources limitations of our computers that could not load large amount of data into memory. An other tricky step was to implement manually the BF algorithm, which do not come out of the box and requires to dig in the `scikit-learn` package Python files to implement the BB from Algorithm 1. Beside these technical difficulties we had issues interpreting our results. First, if our results on HTE can be interpreted through the posterior distribution of parameters we did not understand why the prior information was only used for mathematics concerns and not exploited in a "philosophical way". We thought that it would be the main interest of Bayesian techniques in the case of A/B experiment as marketing and advertisement policies are usually targetting subgroups of consumers defined in advance. Finally, we would have like to make a real comparative study between frequentist and bayesian approach in the case of A/B experiments in particular for trees. Moreover, we would have also liked to test the method proposed in this paper on a usual dataset and compare its performance. We also wanted to apply it on a real dataset but we could not find any that matches our requirements.

Bibliography

[1] Imbens Guido Athey Susan. Recursive partitioning for heterogeneous causal effects. 2016.

[2] Charles J. Stone R.A. Olshen Leo Breiman, Jerome Friedman. *Classification and Regression Trees*. 1984.

[3] Liyun Chen David Draper Matt Taddy, Matt Gardner. A nonparametric bayesian analysis of heterogeneous treatment effects in digital experimentation. 2016.

[4] Donald B. Rubin. The bayesian bootstrap. 1981.

[5] Yu Jun Wyle Mitch Taddy Matt, Chen Chun Sheng. Bayesian and empirical bayesian forests. 2015.