Simulation to check the max purity

2019-03-03

We would like to firstly consider the scenario with two baseline covariates.

We have two treatment arms: placebo (pbo) and drug (drg). The outcomes of those two gruops come from the formula:

$$\mathbf{y} = \mathbf{X}(\beta + \mathbf{b} + \mathbf{\Gamma}(\alpha'\mathbf{x})) + \epsilon.$$

We can define the covariate matrix of $\mathbf{X}$ as $\mathbf{z}$. The $\mathbf{z}$ contains both fixed effects and random effects.

$$\mathbf{z} = \beta + \mathbf{b} + \mathbf{\Gamma}x$$

Parameters:

Two groups

Set parameters:

- $\beta_{drg} = \begin{bmatrix} 0 \\ 25 \\ 1 \end{bmatrix}, \beta_{pbo} = \begin{bmatrix} 1 \\ -5 \\ -1 \end{bmatrix}$
- $\Gamma_{drg} = \begin{bmatrix} 0 \\ -2 \\ -1 \end{bmatrix}, \Gamma_{pbo} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$
- $\mathbf{b} \sim \begin{bmatrix} 1 & 0.1 & 0 \\ 1 & 0.3 & 0 \\ 2 & 0.2 & 0.1 \end{bmatrix} * N(3,1)$
- $\epsilon_{drg} \sim N(3,1); \epsilon_{pbo} \sim N(4,1);$

Baselines

The baselines come from the same distributions

- Baseline covariate x_1, x_2, iid $\sim N(0,1)$
- A true coefficent vector α = (0,1).
- A combination of baseline covariate w: $w = \alpha^T[x_1, x_2]$

Purity calculation

$$p_w(x) = \frac{(f_1(x|w) - f_2(x|w))^2}{f_1(x|w) + f_2(x|w)}$$

where * $f_1(x|w) \sim MVN(\beta_1 + \Gamma_1 * w, \mathbf{b_1})$

- $f_2(x|w) \sim MVN(\beta_2 + \Gamma_2 * w, \mathbf{b_2})$

1. Generate datasets based on the parameters and true α

2. Fit LME and estimate β, Γ and $\mathbf{b}$
3. Calculate the purity based on the above formula

With the true α, the purity should reach the max value.

Then test whether it is correct or not.

1. Choose another α candidate: $\alpha\prime$ and calculate another baseline covariates combination $w\prime$
2. Fit the LME with $w\prime$ and estimate $\beta\prime$, $\Gamma\prime$ and $\mathbf{b}\prime$
3. Calculate the purity based on the above formula

With $\alpha\prime$, the purity should be smaller then the purity calculated by the true α

Results

The purity calculated by true α: 0.5866095

Other α candidates:

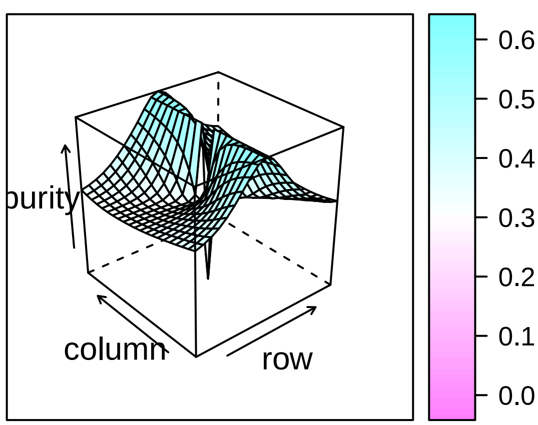
- c(1.1,0): 0.5866102
- c(1,0.5): 0.4853827
- c(1,1): 0.3888147
- c(0,1): 0.3270925
- c(1,10): 0.3310953
- c(-1,1): 0.333868

Find the max

I used the Newton Raphson method to find the max value. However, it still did not work well. We may try some other algorithm.

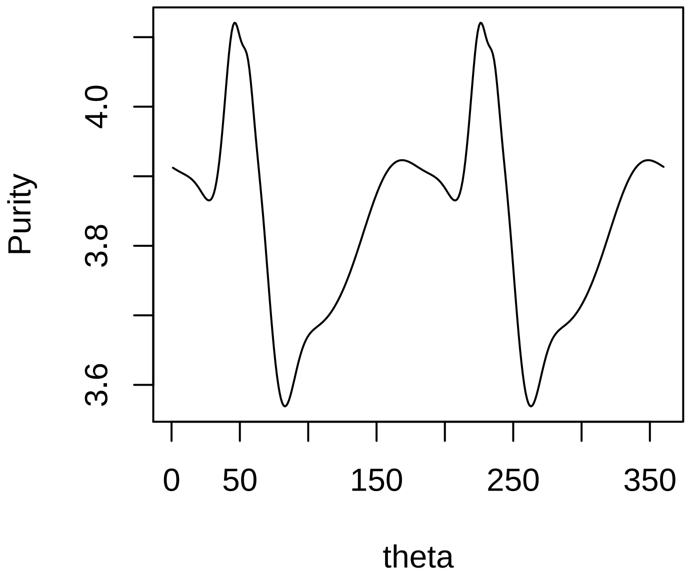
I just simply tried line search method, $\alpha = [\alpha_1, \alpha_2]$, vary α_1 for (-10,10,by =1); vary α_2 for (-10,10,by =1).

The purity looks like:



Purity calculatin 2

- Make the $\alpha = [sin(\theta), cos(\theta)]$. The θ as the only input parameter.
- Set true purity as $\frac{3}{\pi}$



The max purity is:

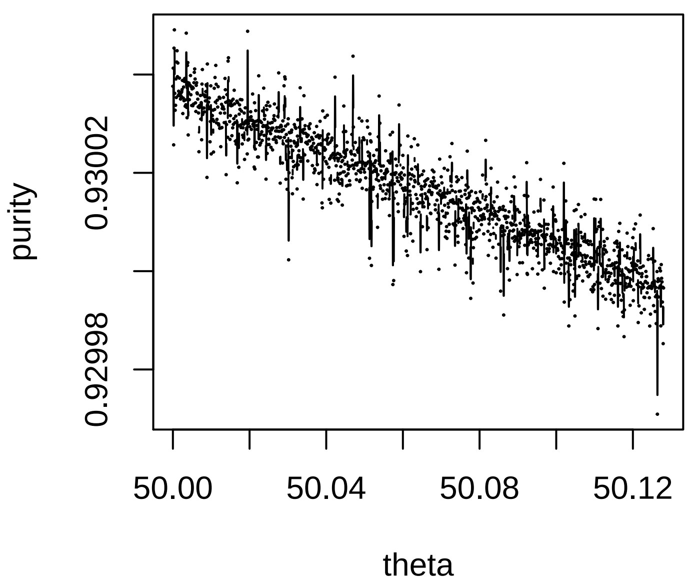
```
data[data$purity == max(data$purity),]
```

```
##     theta   purity
## 46     46 4.120739
## 226   226 4.120739
```

The plot looks smooth. However, it isn't. Since in the previous plot, the distance between two points is 1 degree. Let's make the distance smaller.

Check the points between 50, by 0.0001.

The plots:

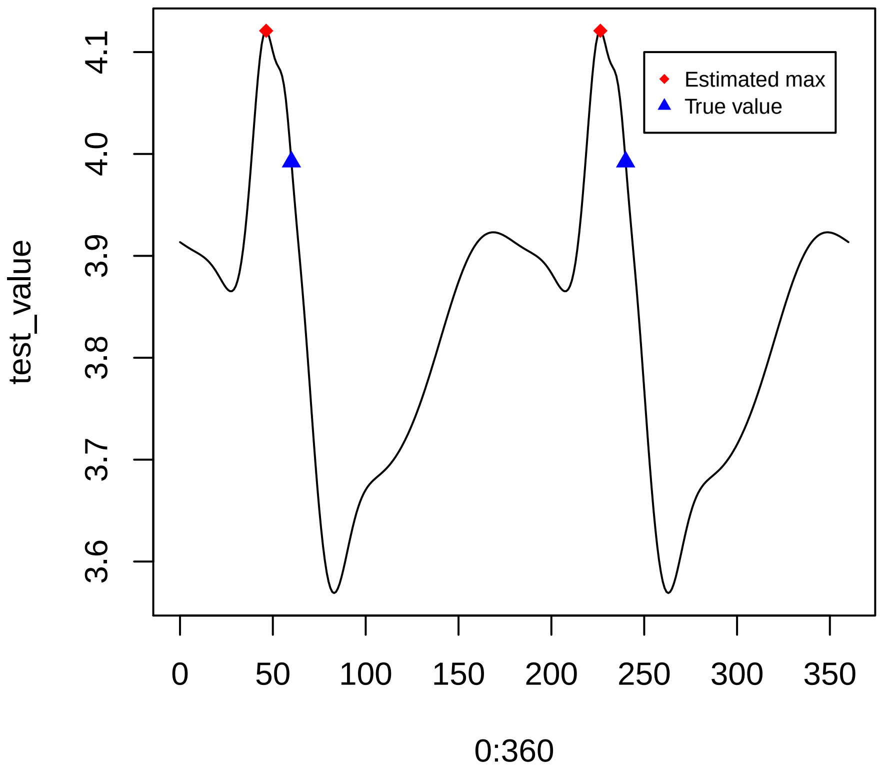


We can see that the plot has a trend, but actually very rough. It is hard to calculate the max value through Newton method.

Global optimization algorithms

There are some algorithms to find the global extreme values instead Newton Raphson method or gradient descent.

Two algroithms: genetic algorithm (GA) and simulated annealing (SA) are commonly used. The genetic algorithm seems to be the most accurate method of the two to find both the maximum and minimum of any function. Therefore, I tried to use GA to find the max value in our purity function. The results is showing below.



The red points is the max value find by the genetic algorithm. We can see this algorithm works well. However, the max value did not match the true value.