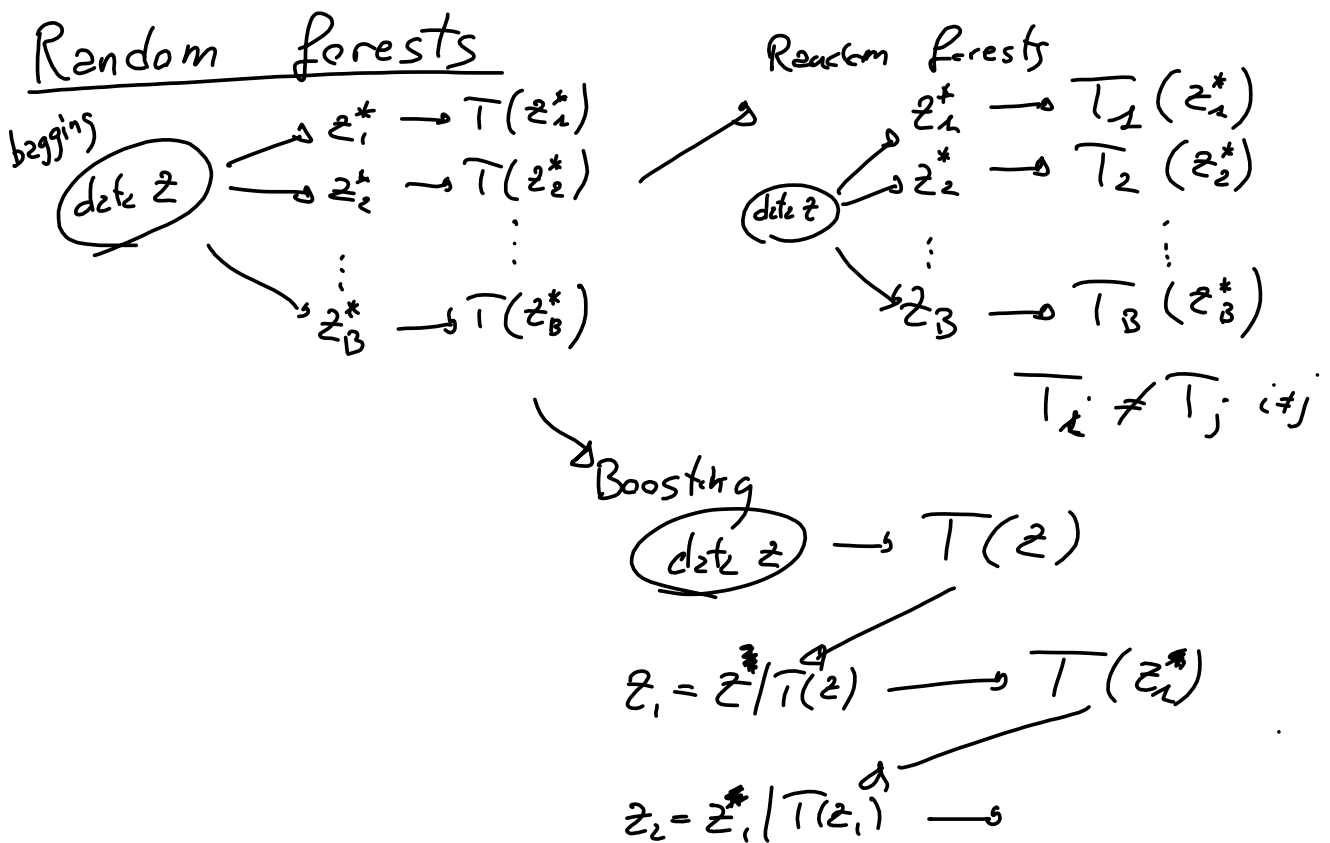




Average of B independent and identically distributed (iid) random variables X with variance σ^2

$\bar{X}$ has variance $\frac{\sigma^2}{B}$

bagging, the trees are not independent.

Average of B identically distributed (id) random variables X with variance σ^2

$$\bar{X} \text{ has variance } \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

where ρ is the correlation between X s

In bagging, we are fitting the same tree on different bootstrap samples $\Rightarrow \rho \neq 0$

By increasing B , we are reducing $\frac{1-\rho}{B}\sigma^2$, but we cannot do anything on $\rho\sigma^2$

Random Forests: Try to reduce the correlation between trees!

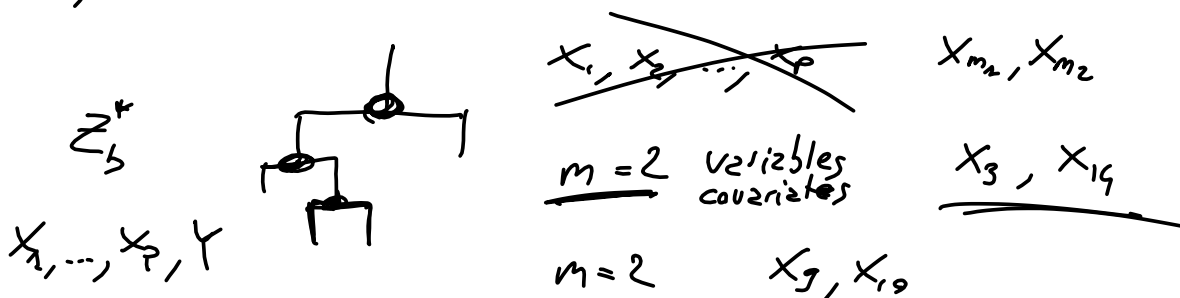
Random forests:

1) For $b = 1, \dots, B$

(a) generate a bootstrap sample (by sampling with replacement from the original sample)

(b) grow a tree in which, at each split, we only consider $m < p$ of the p covariates, chosen randomly

2) average over the B trees (regression)

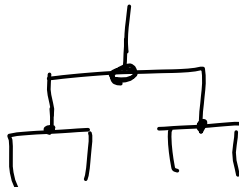


My dataset

$(x_1, x_2, x_3, x_4, x_5, y)$ $p=5$

Bagging

$z_1^* \rightarrow (x_1, \dots, x_5, y)$



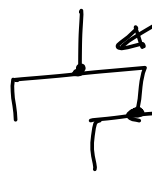
$z_2^* \rightarrow (x_1, \dots, x_5, y)$



I do not consider all $p=5$ covariates, but only $m=2$ of them (e.g., x_2, x_4)

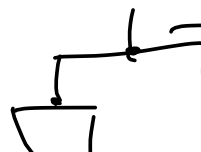
Random forest

$z_1^* \rightarrow (x_1, \dots, x_5, y)$



e.g.; I can only use $m=2$ not necessarily the same as before (e.g., x_2, x_5)

$z_2^* \rightarrow (x_1, \dots, x_5, y)$



e.g. x_1, x_3

The key tuning parameter is m , the fraction of covariates that are picked at each split for possible use.

How to choose m : $\rightarrow CV$
 $\rightarrow$ default value, for regression is $\lfloor P/3 \rfloor$
 for classification is $\lfloor \sqrt{P} \rfloor$

m controls the complexity, i.e., its choice is based on the bias-variance trade-off principle

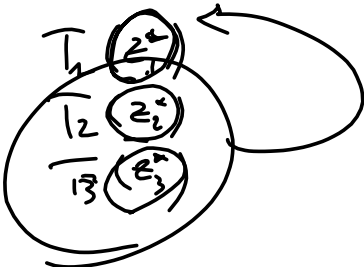
smaller $m \rightarrow$ more complex
 larger $m \rightarrow$ low complexity

Other tuning parameters:

- size of the tree (is not really affecting results)
- size of B (as large as possible, but at a certain point one does not have any improvement)

Out of bag sample

- since we are fitting trees on bootstrap samples, there are trees that do not use some observations. We can use these observations to ~~test~~ compute the error of an estimator only ^{averaging} based on trees that do not use these observations $\rightarrow$ every time we have an estimator of the error



We can measure the variable importance

- fit B random forests with the covariate x_j , and with ~~the covariate~~ x_j with permuted values, and compute the difference (e.g. RSS)

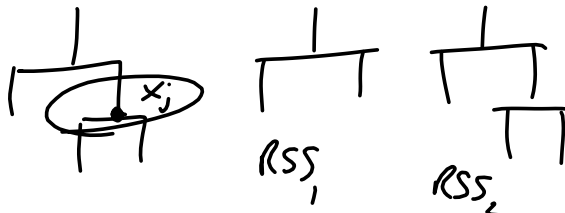
y	x_j
1	2
0	0
-1	2

$$f(x_j) = \frac{1}{2} x_j$$

y	x_j^p
1	-2
0	2
-1	0

$$f(x_j^p) = ? < f(x_j)?$$

- check how much decrease in the loss (RSS) we get at each split that involves a specific variable x_j



each split $\mathcal{I}_e(T) = \sum_{t=1}^{J-1} i_t^2 \mathbb{1}(v(t) = j)$

here $j = l$

Average over all the trees of the forest

$$\mathcal{I}_e^2 = \sum_{m=1}^M \mathcal{I}_e^2(T_m)$$