© Théo Jalabert

Machine Learning, Neural Networks and Deep Learning

Dr. B. Wilbertz[1]

[1]Trendiction S.A. / Talkwalker

15th October, 2021

Model ensembling

© Théo Jalabert

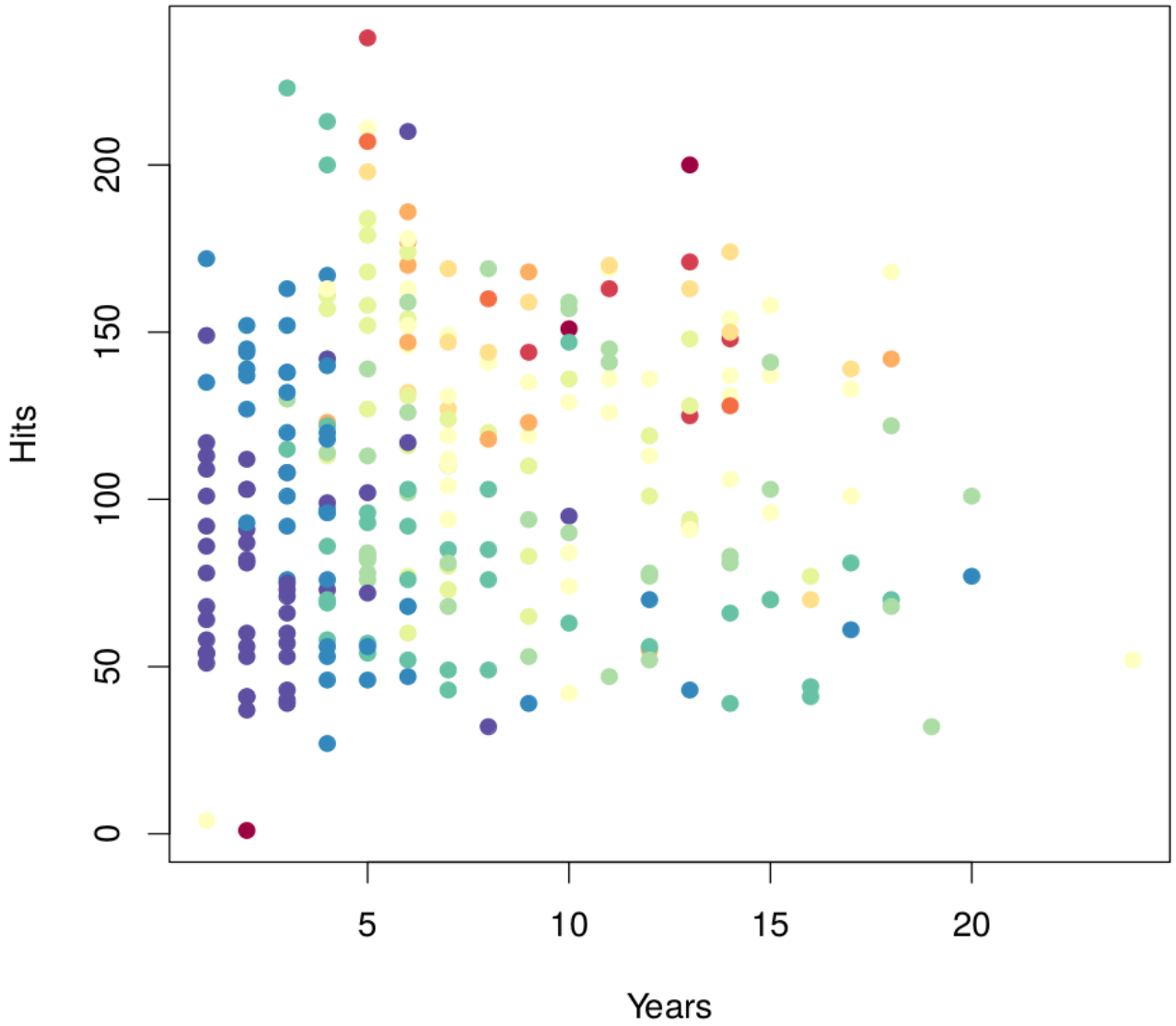
Recall idea of Random Forest

- Average many *weak learners* to form a strong learner
- Given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ of the observations is given by σ^2/n
- Instead of taking the *single best* learner, we trade some bias against reduction in variance

Recall Baseball salary data

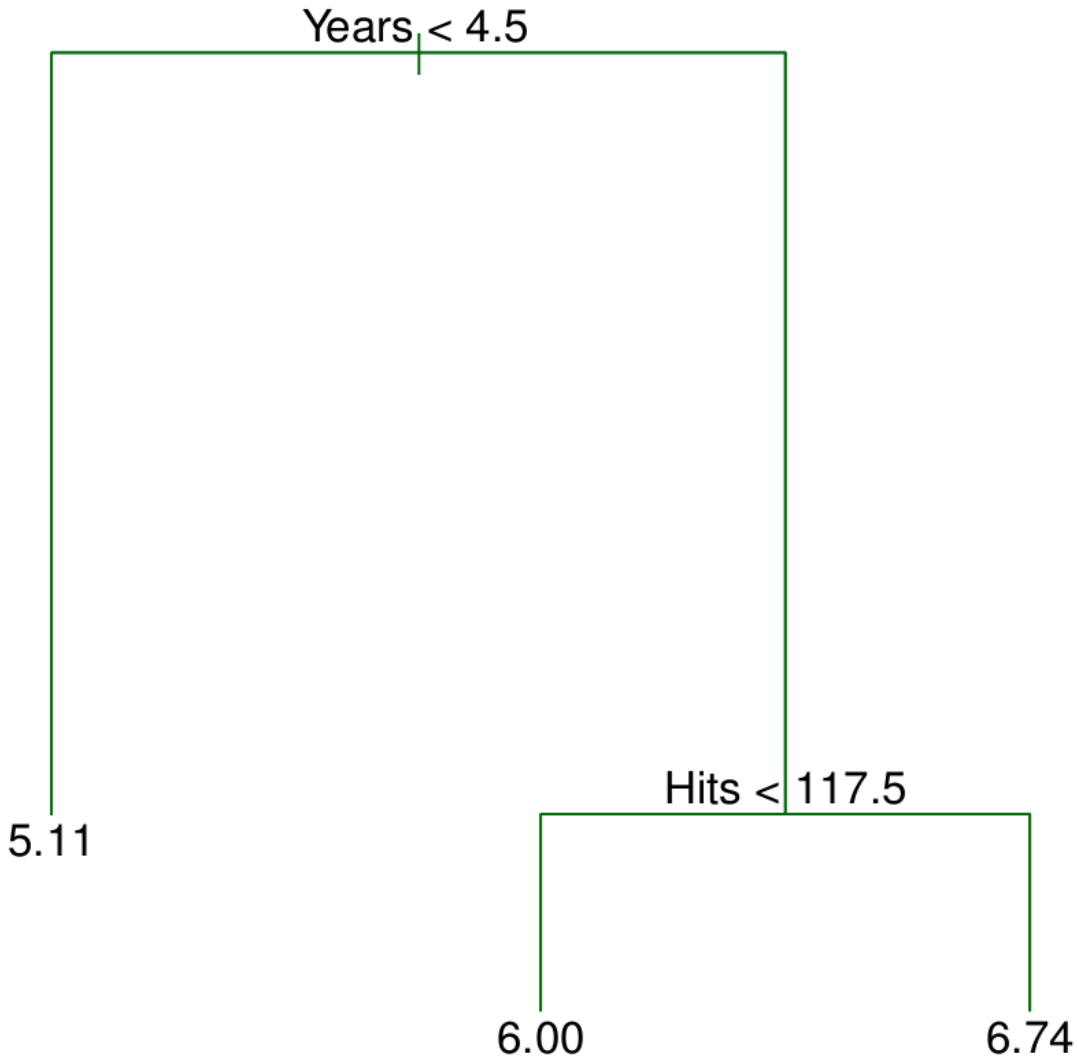
© Théo Jalabert

Salary is color-coded from low (blue, green) to high (yellow,red)



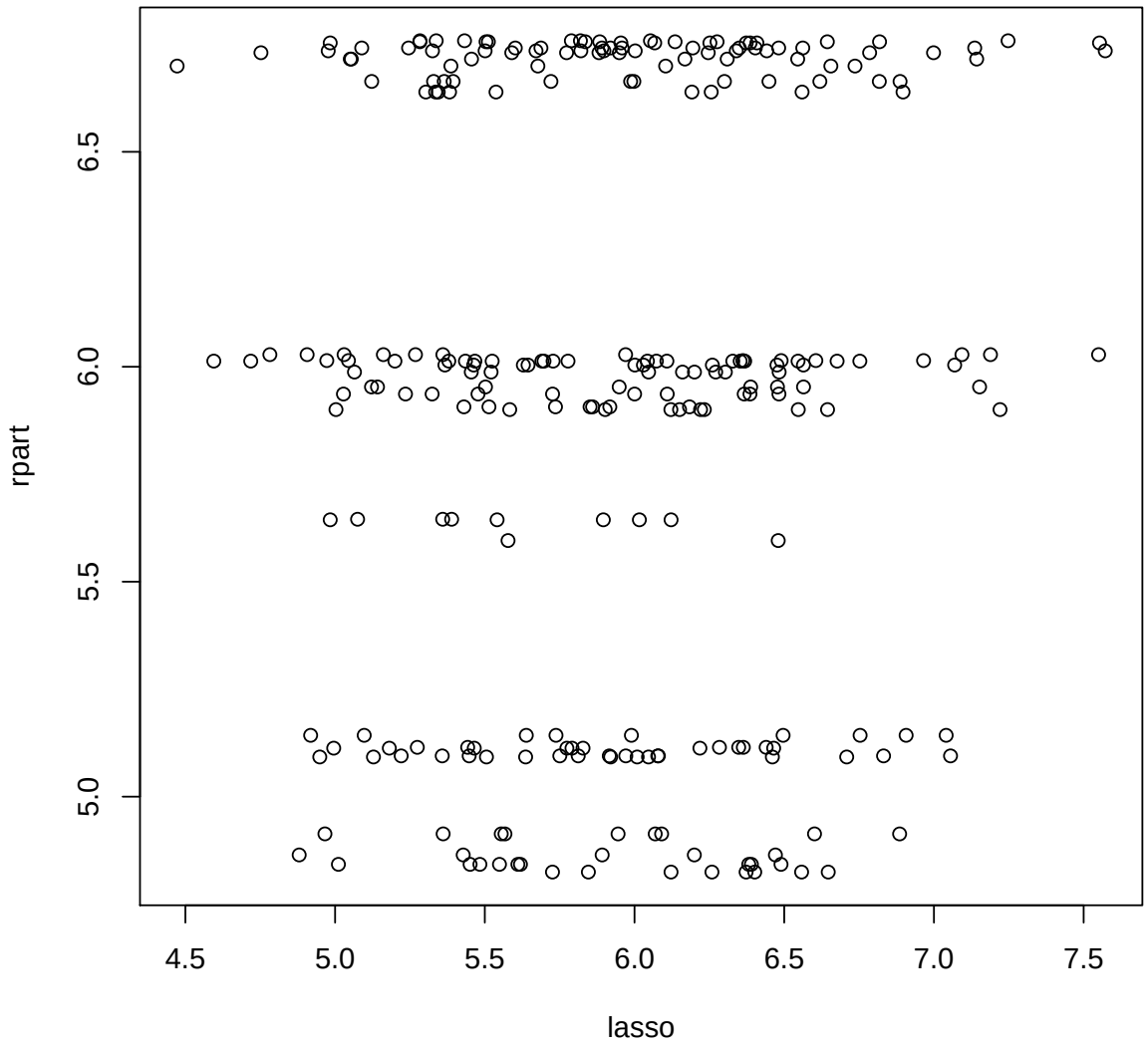
Decision tree for these data

© Théo Jalabert



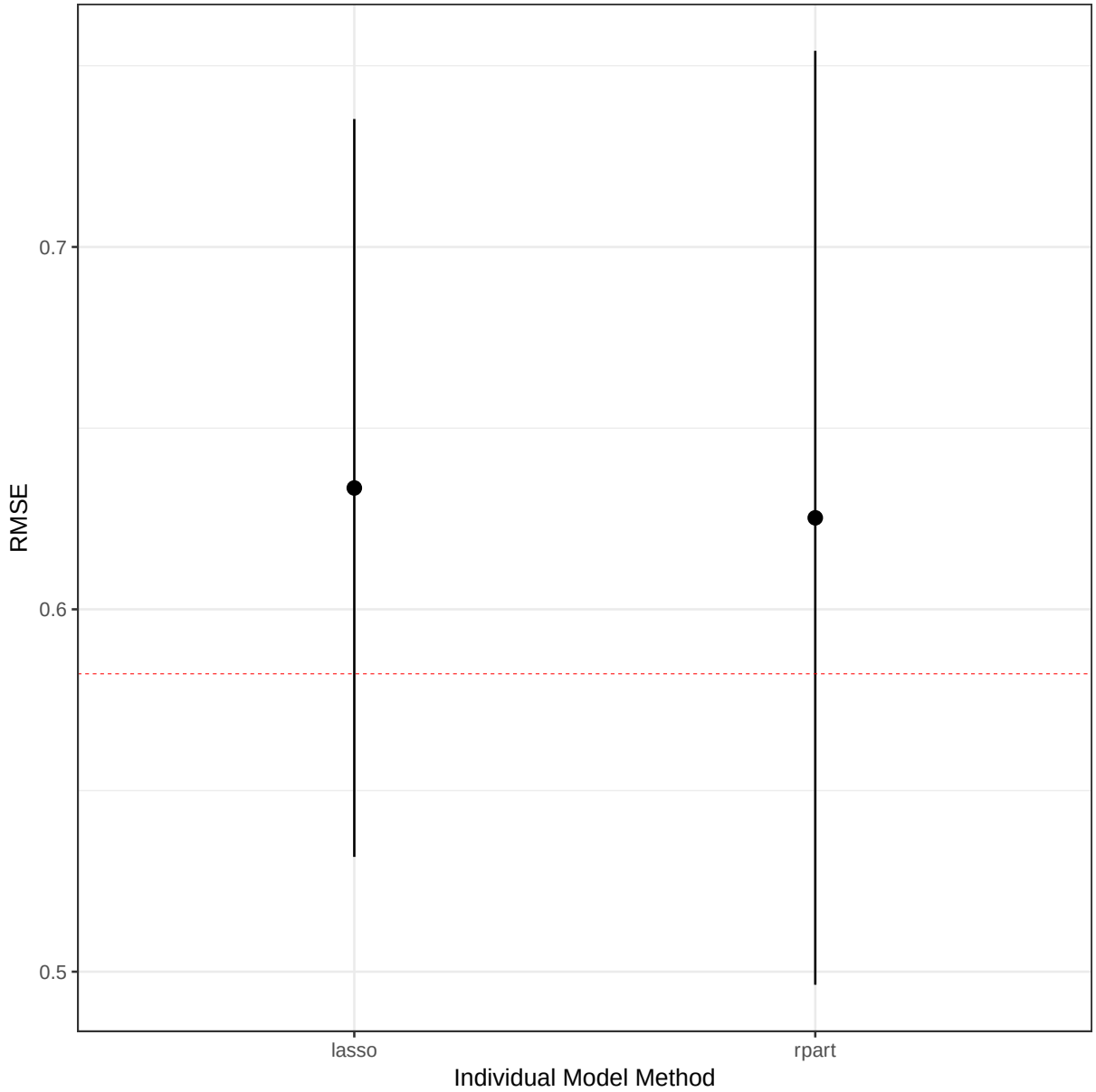
Lasso vs Decision Tree Predictions

©Théo Jalabert



RMSE of Lasso and Decision Tree (10-fold CV)

©Théo Jalabert



Ensembling Lasso and Decision Tree

© Théo Jalabert

Plain vanilla ensembling

- Average the predictions for regression tasks
- Perform a majority vote in case of classification tasks

Weighted ensemble

```
> summary(ensemble)
The following models were ensembled: lasso, rpart
They were weighted:
-0.8763 0.5862 0.5596
The resulting RMSE is: 0.5857
The fit for each individual model on the RMSE is:
 method      RMSE    RMSESD
  lasso 0.6334986 0.1018139
  rpart 0.6252614 0.1288656
```

Estimating ensemble weights

© Théo Jalabert

Cross-validation for ensembles

- Apply the same cross validation partition simultaneous to all ensemble models
- As results we obtain out-of-sample predictions for every model on every fold (i.e. the whole training set)
- On this new training set we chose the optimal least square weights for the ensemble models and evaluate the error by any choice of resampling methods.

Weighted ensemble on the baseball data

@Théo Jalabert

Weighted ensemble on the baseball data

```
> print(ensemble$ens_model, showSD = T)
Generalized Linear Model

263 samples
  2 predictors

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 263, 263, 263, 263, 263, 263, ...
Resampling results (values below are 'mean (sd)'):

  RMSE                      Rsquared                   MAE
  0.5856648 (0.06098665)  0.5773628 (0.08038734)  0.4267806 (0.
```

Model Stacking

© Théo Jalabert

Stacking

- *Model Stacking* extends the idea of the weighted ensemble and uses a *meta learner* to find the best combinations of the ensemble members
- Weighted or stacked models offer in general huge improvements over simple ensemble averaging in case of heterogeneous ensembles
- We can use here all the canonical tricks from previous lecture in order to produce *uncorrelated* models:
  - Bagging (sampling on the observation level)
  - Column/Feature sampling
  - Even boosting itself can be used as stacking technique
- We also can add many models of the same type but with different hyper-parameters to the ensemble

Model Interpretability

© Théo Jalabert

Importance of Model Interpretability

- If a machine learning model performs well, why not just trust the model and ignore why it made a certain decision?
- *The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.*
- Important to understand mechanism driving the outcome

Partial dependence plots

© Théo Jalabert

Idea for partial dependence plots

- The partial dependence plot (PDP or PD plot) shows the marginal effect of a feature on the predicted outcome of a previously fit model
- The prediction function is fixed at a few values of the chosen features and averaged over the other features.
- A partial dependence plot can show if the relationship between the target and a feature is linear, monotonic or more complex

Partial dependence plots

© Théo Jalabert

Definition

Let $S \subset \{1, \ldots, p\}$ be a subset of predictor indices and C its complement. Then the *partial dependence function* of f on x_S is given by

$$f_S(x_S) = \mathbb{E}_{x_C}[f(x_S, x_C)] = \int f(x_S, x_C) d\mathbb{P}(x_C)$$

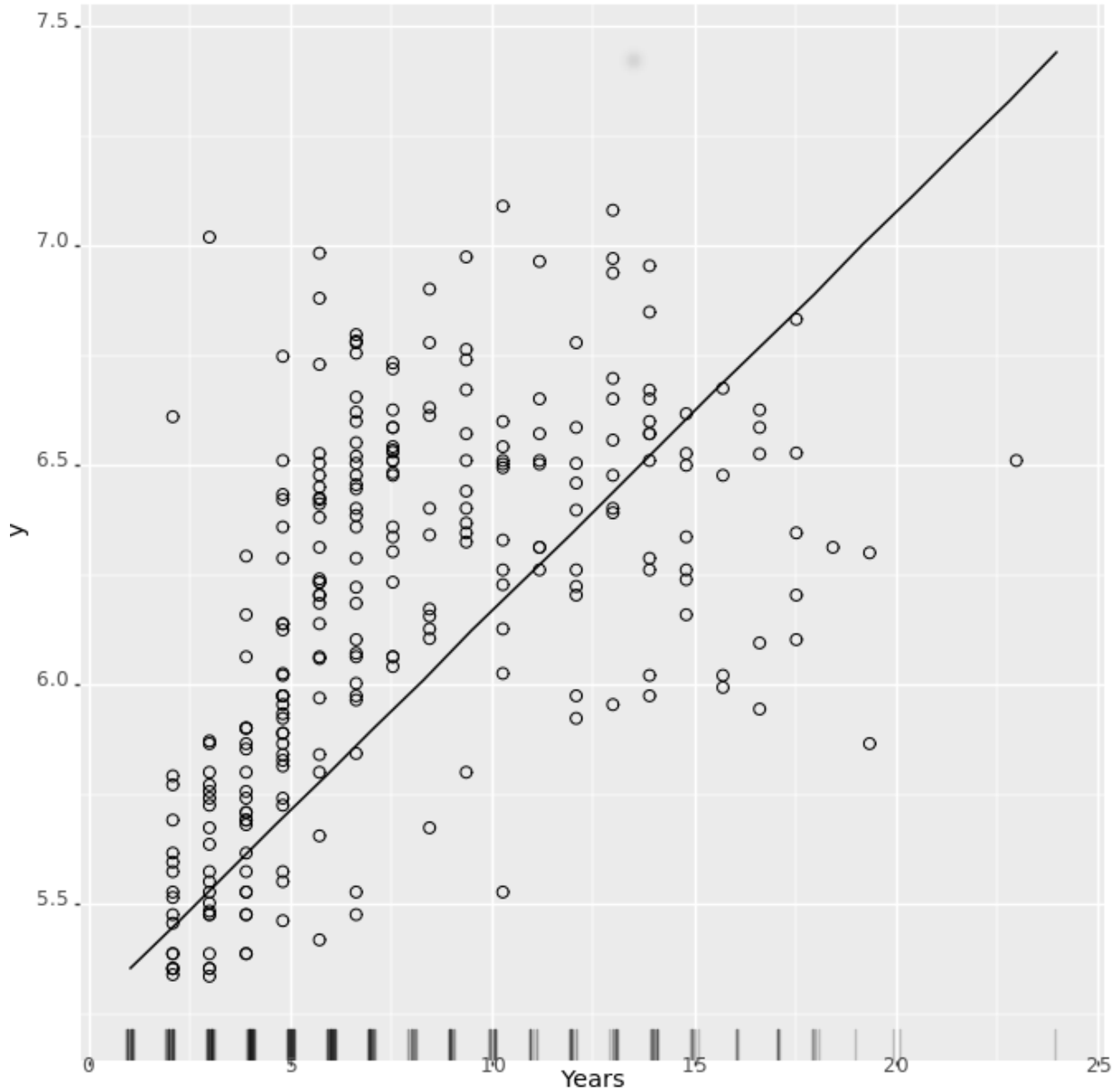
As neither the true f nor $d\mathbb{P}(x_C)$ are known, we estimate f_S by computing

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_{i,C})$$

where $\{x_{1,C}, \ldots, x_{n,C}\}$ represent the different values of x_C that are observed in the training data.

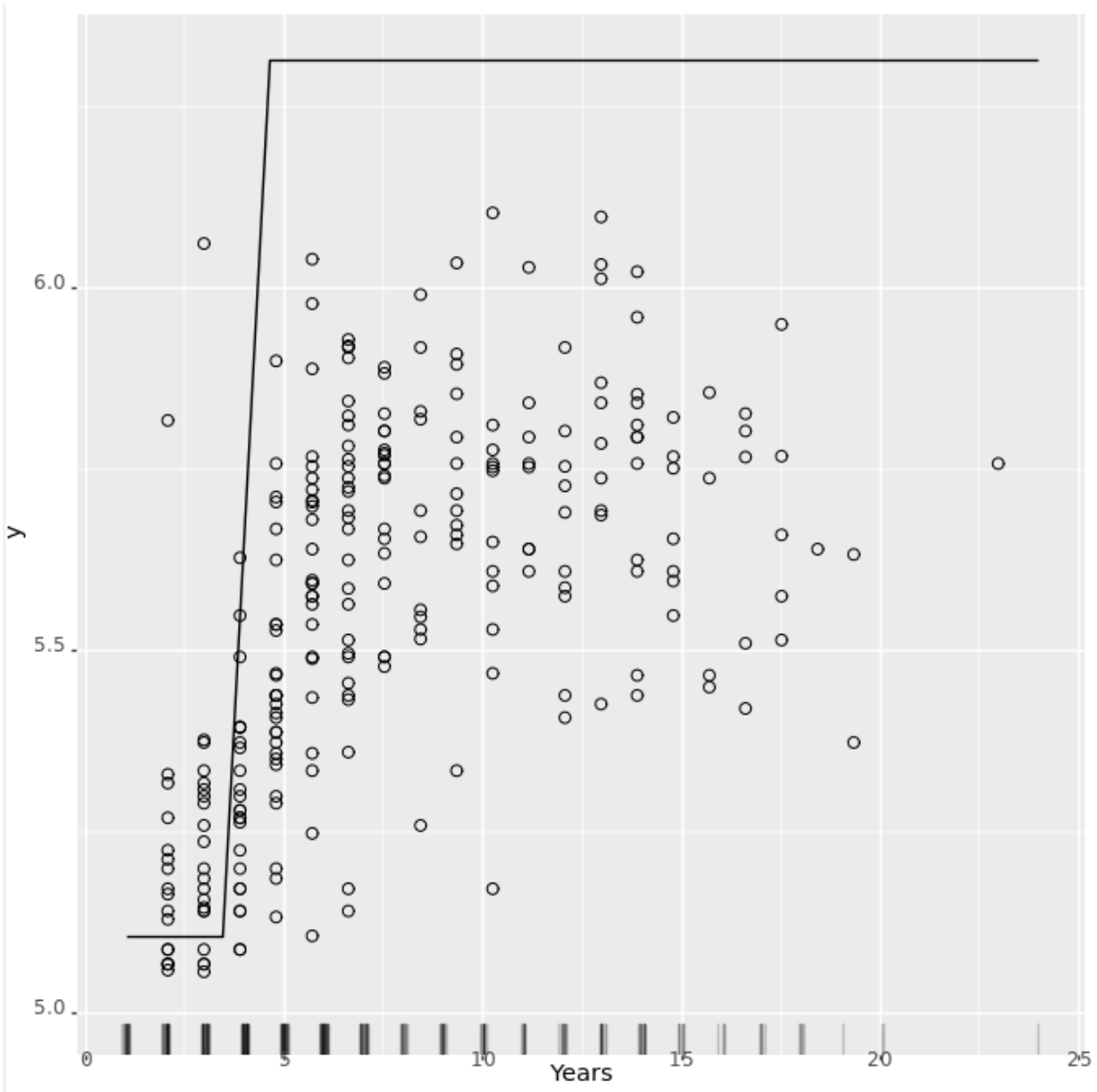
Evaluating $\hat{f}_S$ at every x_S from the training data yields the *partial dependence plot.*

Baseball Example - PDP Lasso © Théo Jalabert



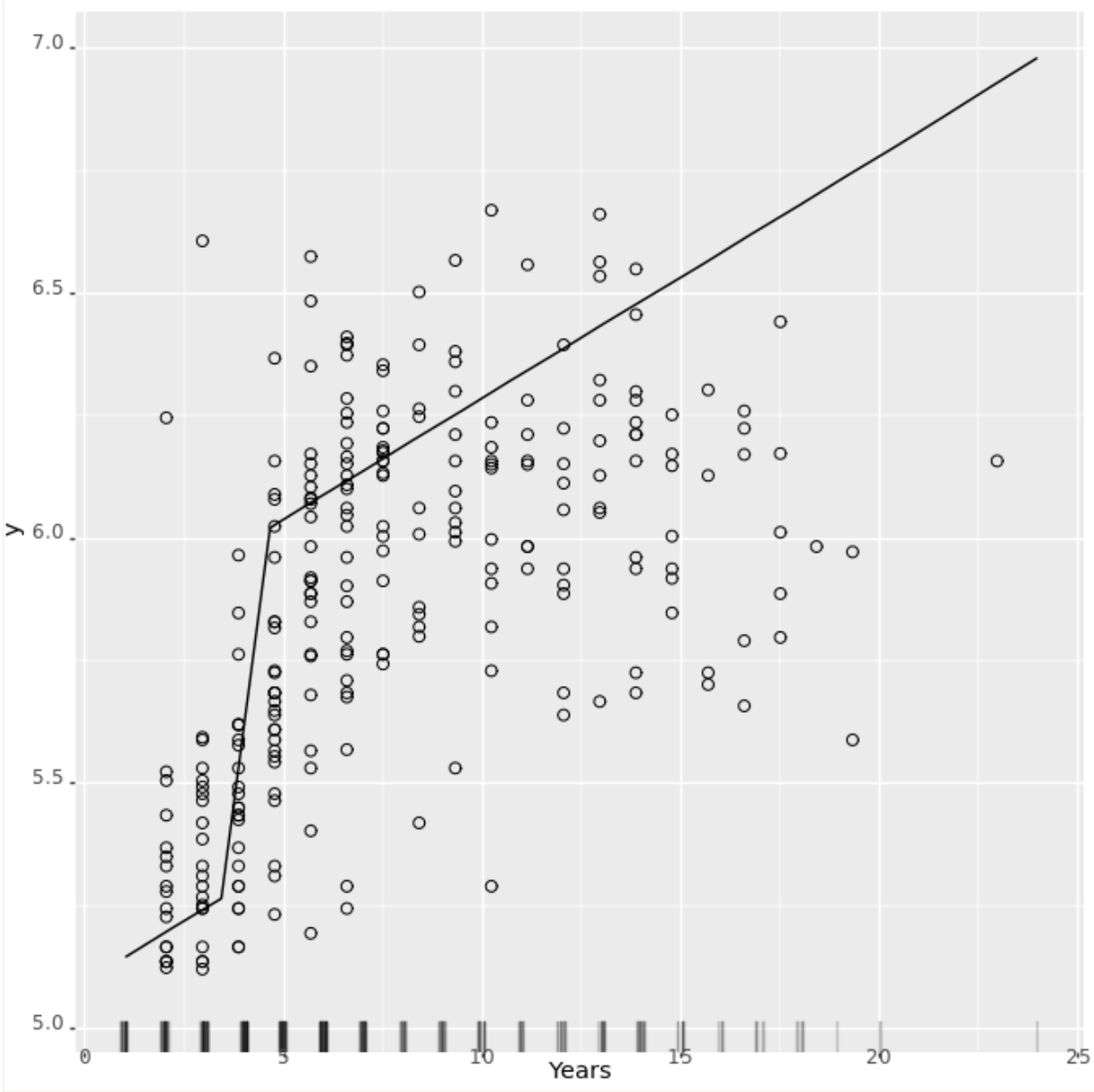
Baseball Example - PDP Decision Tree

©Théo Jalabert



Baseball Example - PDP Ensemble

© Théo Jalabert

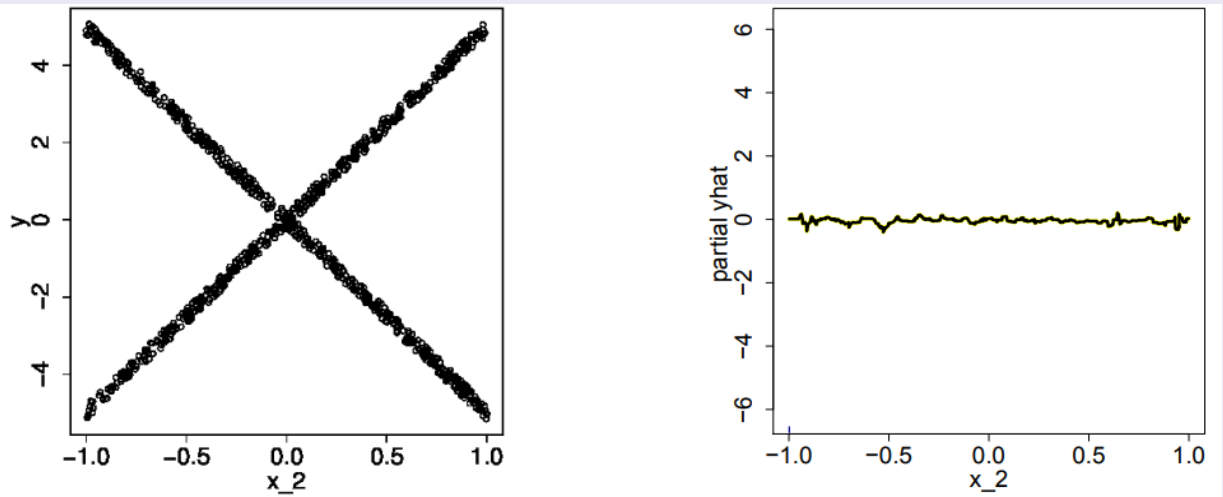


Drawbacks of PDPs

© Théo Jalabert

Drawbacks of PDPs

Partial dependence plots can be misleading when there are interactions to other predictors which average out in $\hat{f}_S$:



Here the PDP incorrectly suggests that there is no meaningful relationship between x_2 and the predicted Y.

ICE Plots

© Théo Jalabert

Idea of ICE plots

Disaggregate the output of classical PDP

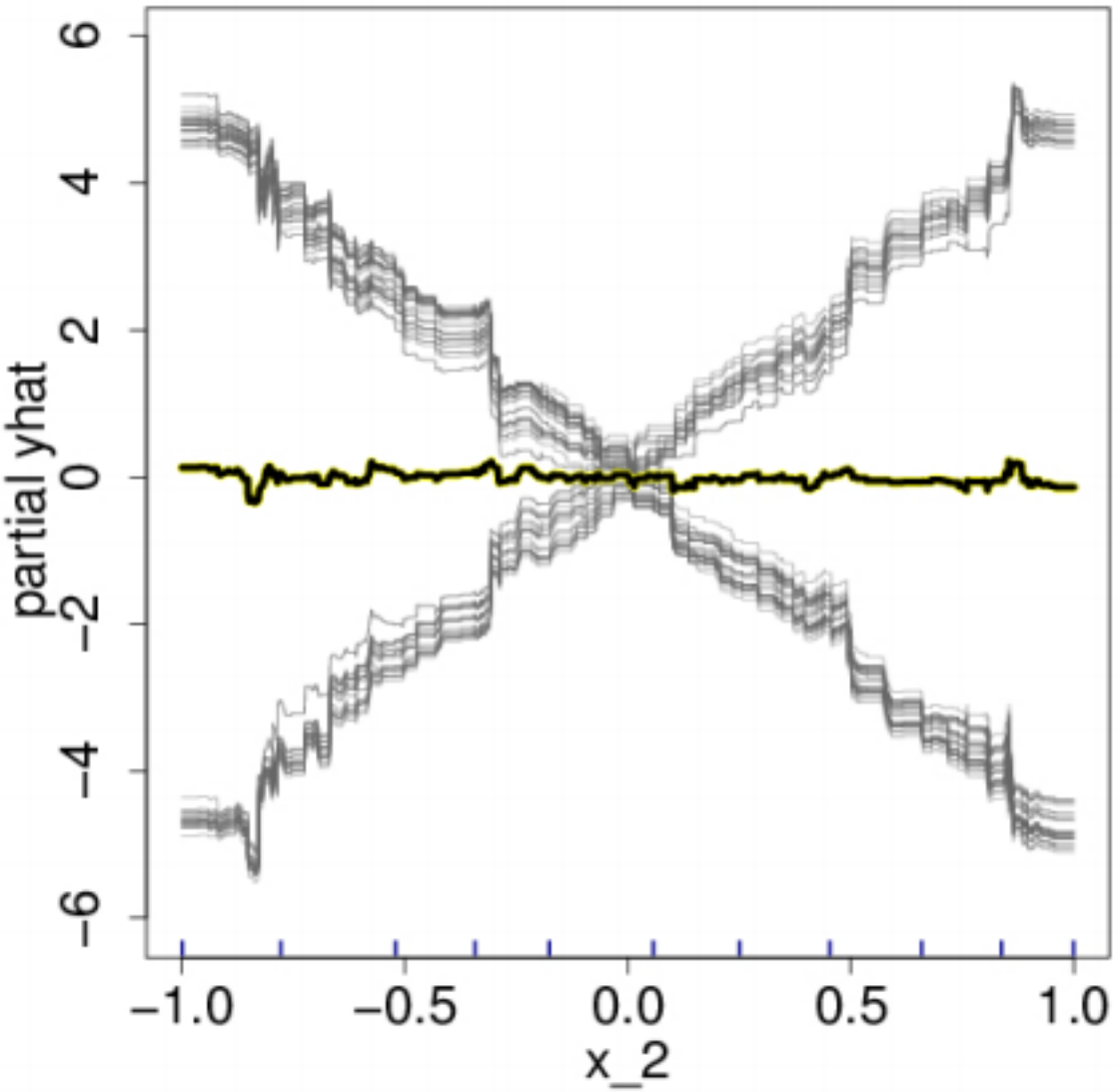
- Rather than plot the target predictors' average partial effect on the predicted response, plot the n estimated conditional expectation curves
- Each curve reflects the predicted response as a function of predictor x_S, conditional on one observed x_C.

Definition

Consider the observations $\{(x_{i,S}, x_{i,C})\}, i = 1, \ldots, n$, and the estimated response function $\hat{f}$. For each of the n observed and fixed values of x_C, a curve $\hat{f}_S^{(i)}$ is plotted against the observed values of x_S.

ICE Plots

© Théo Jalabert



Local (Observation centric) methods

© Théo Jalabert

Local (Observation centric) methods

- Interpretation models that aim at explaining why a single prediction was made that way
- LIME (Local interpretable model-agnostic explanations): fits local, interpretable models in order to explain a prediction
- SHAP (SHapley Additive exPlanations): uses game theory to derive an optimal attribution to features for explaining a predicted outcome

Additive Feature Attribution Methods

© Théo Jalabert

Additive Feature Attribution

- Let $\hat{f}$ be the original prediction model to be explained and g the explanation model.
- Here, we want to explain a prediction $\hat{f}(x)$ based on a single input x.
- Explanation models often use simplified inputs x' that map to the original inputs through a mapping function $x = h_x(x')$ for $x' = \mathbf{1}$.
- Local methods try to ensure $g(z') \approx \hat{f}(h_x(x'))$ whenever $z' = x'$.
- Any model g that is a linear function of binary variables

$$g(z') = \phi_0 + \sum_{i=1}^{m} \phi_i z'_i$$

where $z' \in \{0,1\}^m$, m is the number of simplified input features, and $\phi_i \in \mathbb{R}$, is called *Additive feature attribution model.*

Local interpretable model-agnostic explanations (LIME)

©TheoJalabert

- Let $\mathcal{G}$ be a class of potentially *interpretable* models, such as the additive feature attribution models
- $z' \in \{0,1\}^m$ acts over absence/presence of the *interpretable components*
- As not every $g \in \mathcal{G}$ may be simple enough to be interpretable - we let $\Omega(g)$ be a measure of complexity (as opposed to interpretability) of the explanation $g \in \mathcal{G}$
- We further use $\pi_x(z)$ as a proximity measure between an instance z to x, so as to define locality around x.
- Finally, let $\mathcal{L}(\hat{f}, g, \pi_x)$ be a measure of how unfaithful g is in approximating $\hat{f}$ in the locality defined by π_x.
- LIME is defined as

$$\xi(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathcal{L}(\hat{f}, g, \pi_x) + \Omega(g)$$

Local interpretable model-agnostic explanations (LIME)

©TheoJalabert

- For linear g, one typically choses

$$\mathcal{L}(\hat{f}, g, \pi_x) = \sum_{(z,z') \in \mathcal{Z}} \pi_x(z)(\hat{f}(z) - g(z'))^2$$

and

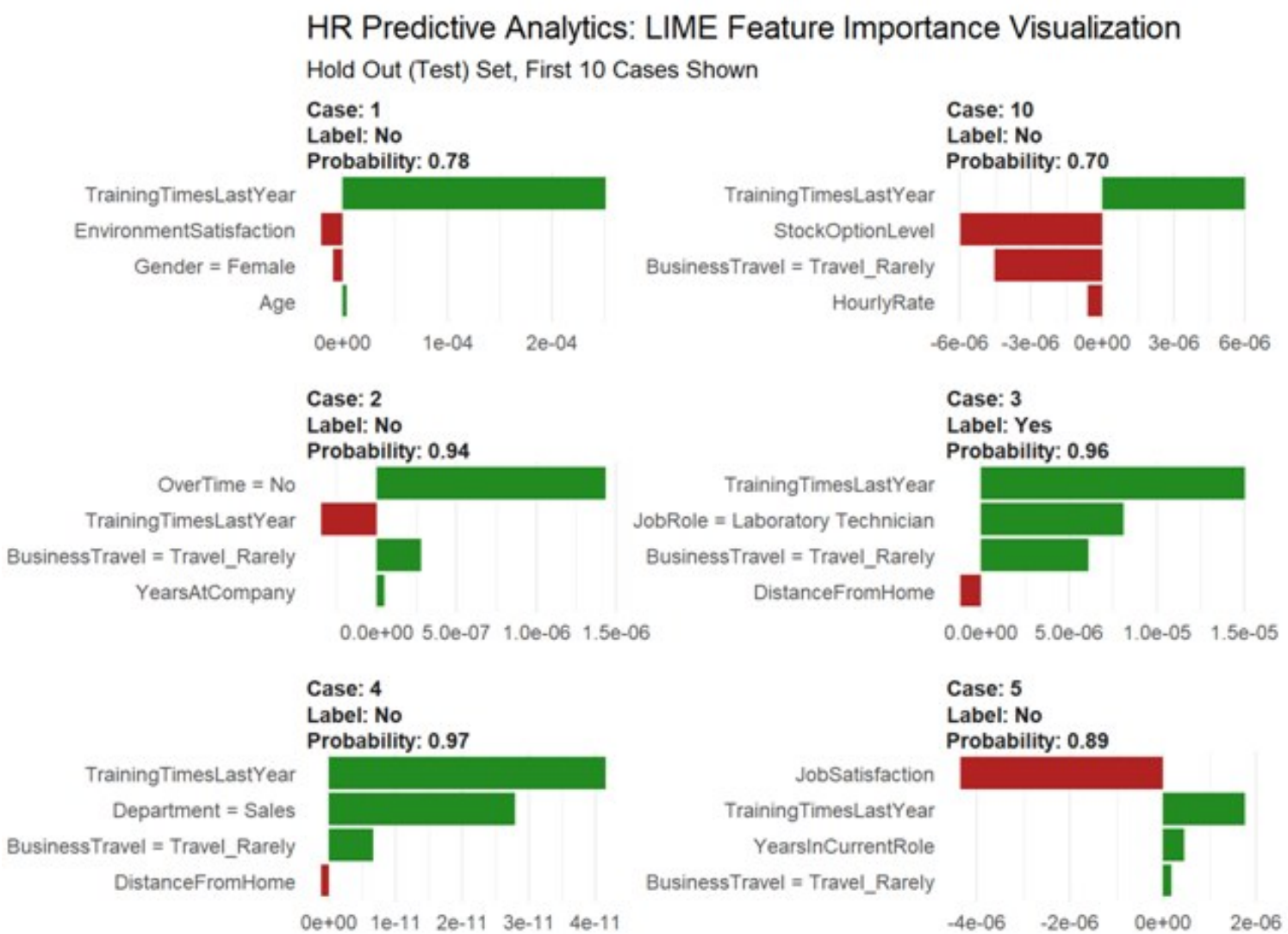
$$\pi_x(z) = \exp(-D(x, z)^2/\sigma^2)$$

for some distance function D and width σ.

- One then samples tuples (z, z') from the neighborhood of x to get a "new training set", such that the optimal g can be found by using penalized regression

LIME on Classification task

© Théo Jalabert



LIME for images

© Théo Jalabert

(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

SHAP (SHapley Additive exPlanations)

© Théo Jalabert

Idea of SHAP

- Find a model g, which fulfills three natural properties of an explanation:
  1. *Local accuracy:* For $x = h_x(x')$, it holds

$$\hat{f}(x) = g(x') = \phi_0 + \sum_{i=1}^{m} \phi_i x'_i$$

  2. *Missingness:*

$$x'_i = 0 \Longrightarrow \phi_i = 0$$

  3. *Consistency:* Let $f_x(z') = \hat{f}(h_x(z'))$ and $z' \setminus i$ denote setting $z'_i = 0$. For any models f, f' and all inputs $z' \in \{0,1\}^d$

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i)$$

     implies

$$\phi_i(f', x) \geq \phi_i(f, x).$$

Shapley values and game theory

© Théo Jalabert

- Consider a game with N players
- Each player has a set of skills, which interact with skills from all the other players
- Question: Given a coalition S of players and a payoff $v(S)$ for this coalition, what is the "fair" distribution of this payoff to the individual players in the coalition?
- A distribution scheme is called "fair" if it satisfies the 3 properties from previous slide
- The only "fair" additive feature attribution model is given

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)],$$

  where N is the set of all features
- SHAP: Consider $f_x(S) = \hat{f}(h_x(z')) = \mathbb{E}[\hat{f}(x)|x_S]$, where S is the set of non-zero indexes in z'.

Computation of SHAP values

© Théo Jalabert

Computation of SHAP values

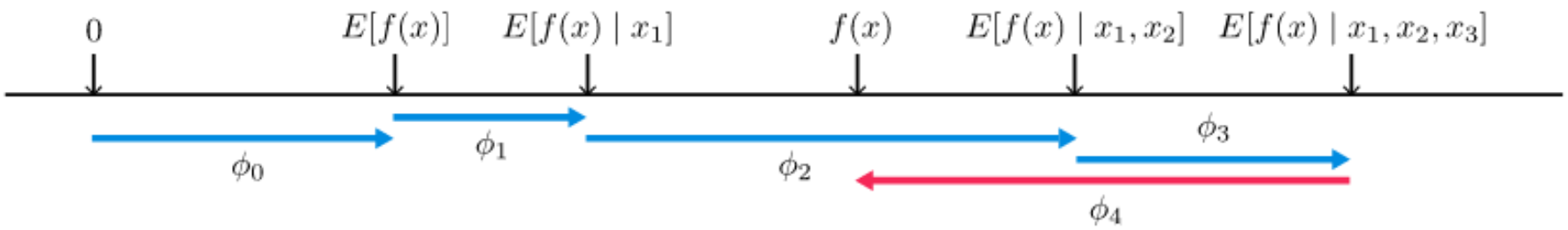
- Unfortunately this formulae involves enumerating all subset of N and therefore grows exponentially
- In the case of a linear model $\hat{f}(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$, we can derive the feature attributions for observation j as

$$\phi_i(\hat{f}, x, j) = \beta_i(x_i^{(j)} - \mathbb{E}[x_i]) \qquad \text{and} \qquad \phi_0(\hat{f}, x) = \beta_0.$$

- In case of decision tree, there also exists an optimized solution which is *only* quadratic in the depth of the tree.
- The optimized version is implemented in the standard gradient boosting packages `XGBoost`, `LightGBM` and `CatBoost`.

Local Accuracy of SHAP explanation / SHAP plots

© Theo Jalabert

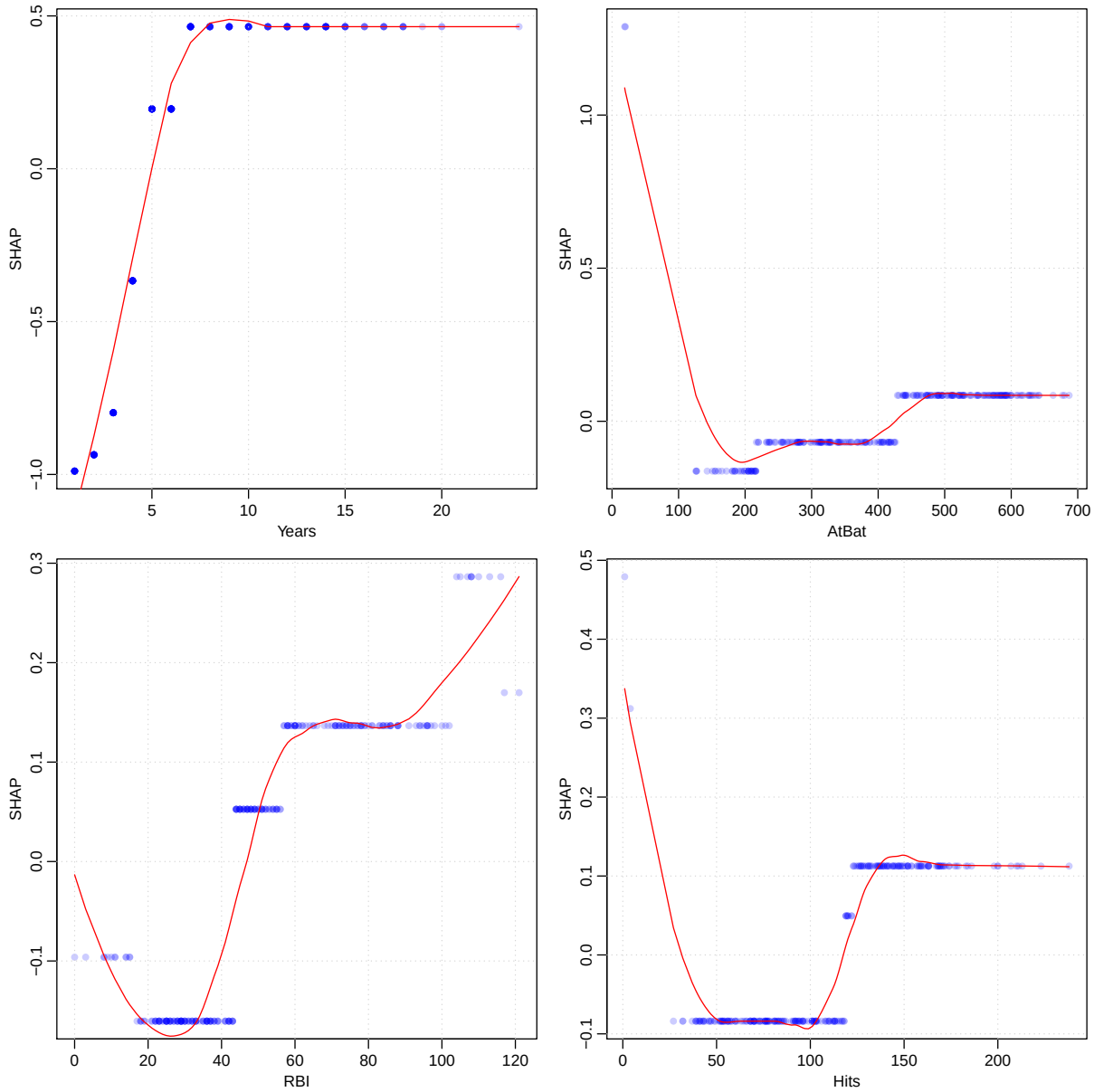


SHAP plots

- In order to get a global model interpretation using SHAP values, we can plot the SHAP values of every data point in our training set againt the underlying feature value.
- Usually, one also fits a *LOESS* curve through these points in order to highlight the general trend

SHAP Baseball example on XGBoost fit

© Théo Jalabert



Common pitfalls when interpreting models

© Theo Jalabert

Common pitfalls when interpreting models

- It is very tempting to ask questions like "How much can I change my outcome Y, when increasing feature i by amount x?"
- Unfortunately, this is a *causal* question
- Machine Learning models only capture *associations* (and this with the only goal to minimize prediction error)
- We can only ask questions like "How much will outcome Y change, when I observe that feature i changes by amount x?"
- Apart from the causality-vs-correlation problem, we also have no guarantee on statistical error bounds for the feature attributions.

Causal inference

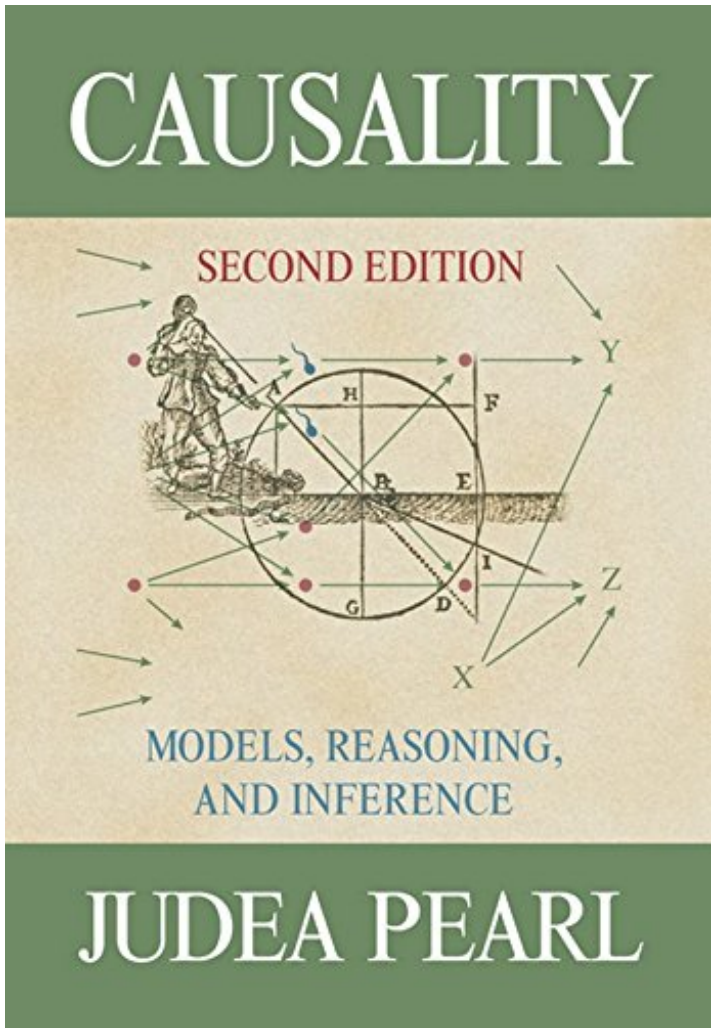
© Théo Jalabert

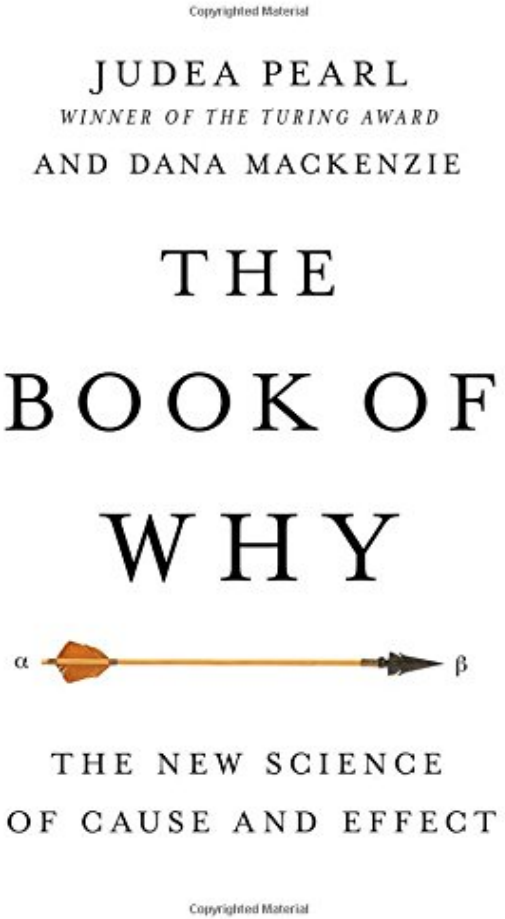
Causal inference

- Process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
- Gold-Standard for questions of causality are *Randomized Controlled Trials* (RCTs)
- But, in many situations RCTs are not possible, ethically unacceptable, too expensive, or just do not yield large enough samples sizes to estimate effects
- Modern causal inference attempts to draw causal conclusion from observational data

Judea Pearl

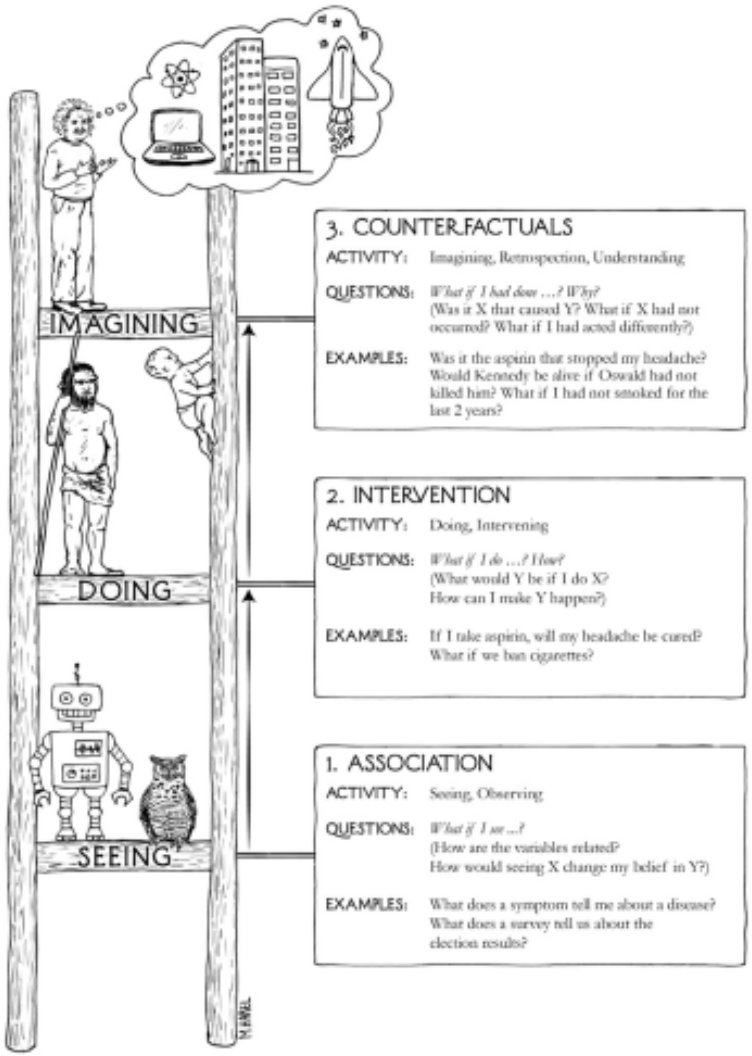
© Théo Jalabert





Judea Pearl's ladder of causality

© Théo Jalabert



Historical problems of causality and statistics

©Théo Jalabert

Causal inference is difficult...

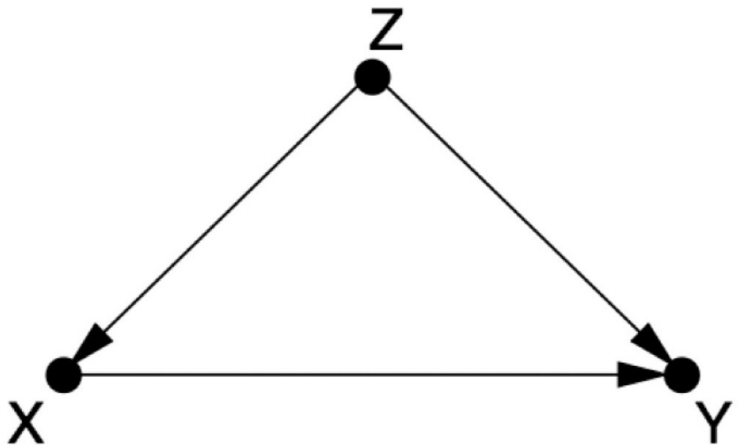
- mathematical language not adequate to formulate causal relations
- wrong definition of causal effect (X causes Y, when X increases $\mathbb{P}(Y|X)$)
- misleading concept of correlation (Galton/Pearson): Unable to explain spurious correlation

Example (Spurios correlation)

- Chocolate consume per capita and nobel prizes are positively correlated over countries.
- But, eating more chocolate will hardly produce more nobel prizes
- Correct explanation: Countries, which consume more chocolate are usually wealthier countries, and those have also better education and higher chances for winning nobel prizes

Causal diagrams / DAGs

© Théo Jalabert



Causal diagram

- The key into causal inference is to precisely define the relations of the variables from our datasets within a causal diagram.
- $X \to Y$ means, Y is "listening" to X, i.e. it determines its value in response to what it hears from X.

Causal diagrams / DAGs

© Théo Jalabert

Main components of causal DAGs

- A→B→C "Chain" B is a mechanism that transmits the effect of A to C. Moreover, B "screens off" information about A from C, i.e. once we know about B the effect on C is independent of A
  Example: Fire → Smoke → Alarm. We need to observe smoke for the alarm to trigger. But when we have smoke, the alarm will trigger, independently of the smoke being caused by a fire or something else.
- A←B→C "Fork" and B is often called a confounder of A and C, i.e. it makes A and C statistically correlated even though there is no direct causal link between them. As with the chains, A and C are conditionally independent, given B.
  Example: Shoes size ← Age of Child → Reading Ability. Children with larger shoes tend to read at higher level. But the relation is not one of cause and effect. Giving a child larger shoes won't make him read better. (Spurious correlation)

Causal diagrams / DAGs

© Théo Jalabert

Main components of causal DAGs

- A→B←C "Collider", that is if A and C are independent in the beginning, conditioning on B will make them dependent.

  Example (with Hollywood actors): Talent → Celebrity ← Beauty. We assert that talent and beauty contribute to actor's success, but they are completely unrelated to one another in general population. If we look only at famous actors (i.e. Celebrity = 1), we will see a negative correlation between talent and beauty: finding out that a celebrity is unattractive increases our belief that he or she is talented.

Confounding

© Théo Jalabert

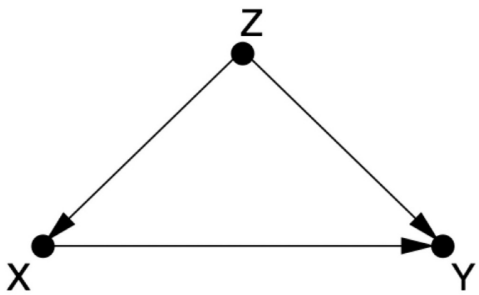


FIGURE 4.1. The most basic version of confounding: Z is a confounder of the proposed causal relationship between X and Y.

Confounding

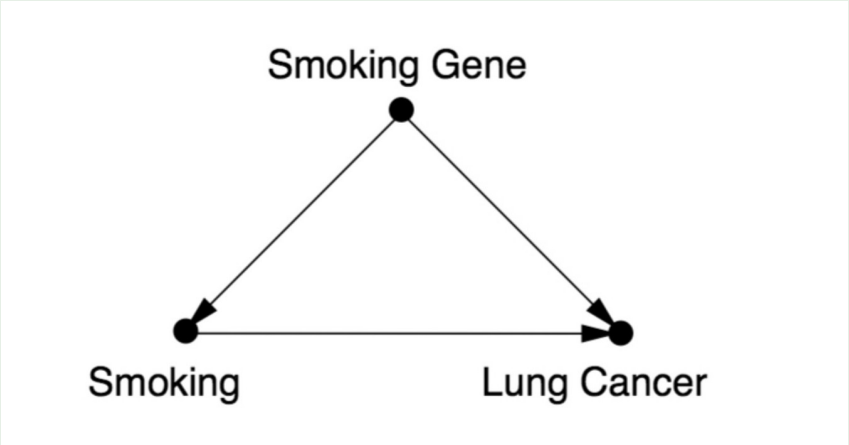
- The term "confounding" originally meant "mixing" in English and we can understand from above diagram why the name was chosen.
- The true causal effect $X \to Y$ is "mixed" with the spurious correlation between X and Y induced by the fork $X \leftarrow Z \to Y$.
  Example: If we are testing a drug and give it to patients who are younger on average than the people in the control group, then age becomes a confounder – a lurking third variable. If we don't have any data on the ages, we will not be able to disentangle the true effect from the spurious one.

Confounding

© Théo Jalabert

Confounding is a serious issue in estimation of causal effect from observational data.

Example (Is smoking causing cancer?)



- Due to missing data for that potential smoking gene, it was impossible to answer this question from observational studies
- Solution at this time was sensitivity analysis, i.e. assuming that such an unobserved confounder like the smoking gene exists and estimate from data how strong this must be correlated with smoking and lung cancer.

RCTs

© Théo Jalabert

Causal effect estimation from R.A. Fisher

- Classical answer from statistics for investigation of causal question: RCT (Randomized controlled trial)
- Example: Does a new drug X cures a disease D?
- In experiments it is very hard to systematically rule out all possible side effects on drug X's influence on D (gender, genetics, etc), Fisher's idea was instead of stratifying on all confounders, to randomly assign which patients should get the new drug and which ones not (control group).
- Since in RCTs assignment is completely independent of all other variables, confounders cannot act on the result in a systematic way anymore.
- RCT are still gold standard for estimation of causal effect

do-Operator

© Théo Jalabert

Definition of the *do*-Operator

- Pearl's way of mathematically describing this randomized intervention from RCTs is the *do*-operator.
- $do(X = x)$ means that we force X to be x by means of an external intervention (i.e. we play God and role our own the dices)
- In the language of causal diagrams, this means that we delete all incoming paths into X

- Generally, we have $\mathbb{P}(Y|do(X)) \neq \mathbb{P}(Y|X)$ (this means: Y, when we force X to be x, is different from Y when we observe that X equals x)
- We call everything which makes $\mathbb{P}(Y|do(X))$ differ from $\mathbb{P}(Y|X)$ confounding.

Causal effect and Backdoor paths

©Théo Jalabert

Goal: Estimate causal effect and establish methodology to find "correct" set of variables to control for confounding in observational data

Remember: *do*-Operator erases all incoming arrows, and prevents information from flowing in non-causal direction (i.e. emulates a RCT)

So far we have seen three rules, how to stop ("*block*") flow of information:

a. In a chain, A→B→C, controlling for B prevents information about A from getting to C or vice versa.
b. In a fork, A←B→C, controlling for B prevents information about A from getting to C or vice versa.
c. In a collider, A→B←C, exactly the opposite holds: A and C start out independent (information about A tells you nothing about C), but if you control for B, then information starts flowing from A to C or vice versa.

Backdoor criterion

© Théo Jalabert

Definition (Backdoor Criterion)

A set of covariates Z satisfies the *backdoor* criterion relative to variables (X, Y) in a DAG G if:

1. no node in Z is a desendant of X
2. Z blocks every path between X and Y that contains an arrow pointing into X.

Theorem

If a set of covariates Z satisfies the backdoor criterion relative to (X, Y), then the causal effect of X on Y is identifiable and it holds

$$\mathbb{P}(Y|do(X), Z) = \mathbb{P}(Y|X, Z).$$

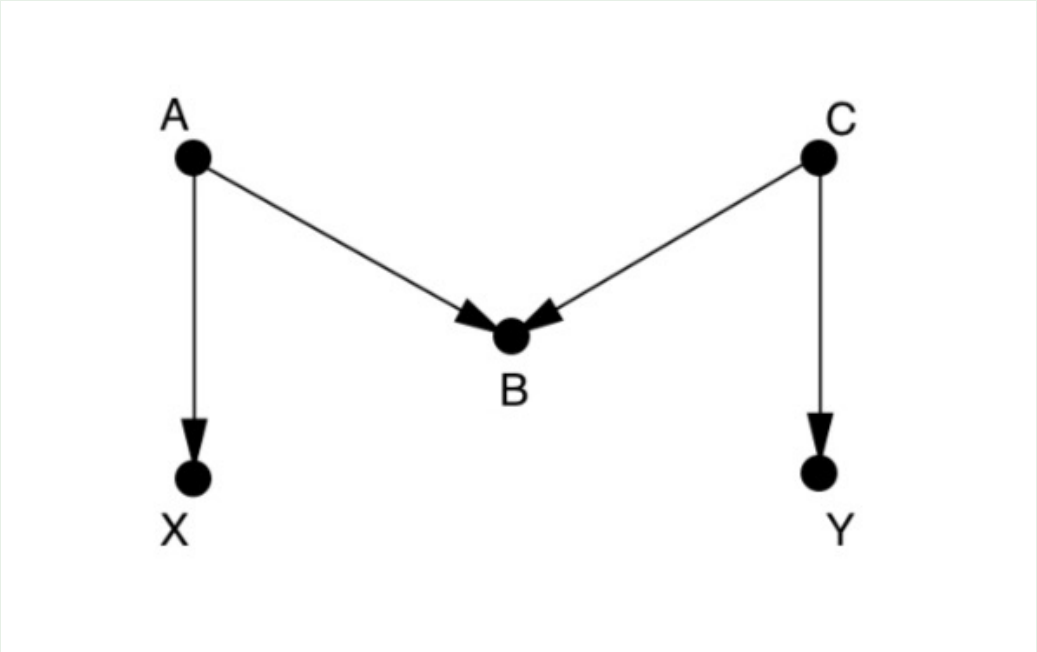
Consequently, we can computed for above Z the causal effect

$\mathbb{P}(Y|do(X = 1), Z) - \mathbb{P}(Y|do(X = 0), Z)$ as
$\mathbb{P}(Y|X = 1, Z) - \mathbb{P}(Y|X = 0, Z)$

Backdoor criterion

© Théo Jalabert

Example (M-Bias)



More backdoor examples

© Théo Jalabert

Birth-weight paradox

- Study on 15000 children in San Francisco Bay Area
- Contains information on
  - mothers' smoking habits
  - birth weights of babies
  - mortality rates of babies in the first month of life
- Expected:
  - smoking has an effect on birth-weight
  - birth-weight has an effect on survival rate
- Findings:
  - babies of smokers were lighter on average than the babies of nonsmokers
  - low-birth-weight babies of smoking mothers had a better survival rate than those of nonsmokers

Question: Has mother's smoking a protective effect?

Birth-weight paradox

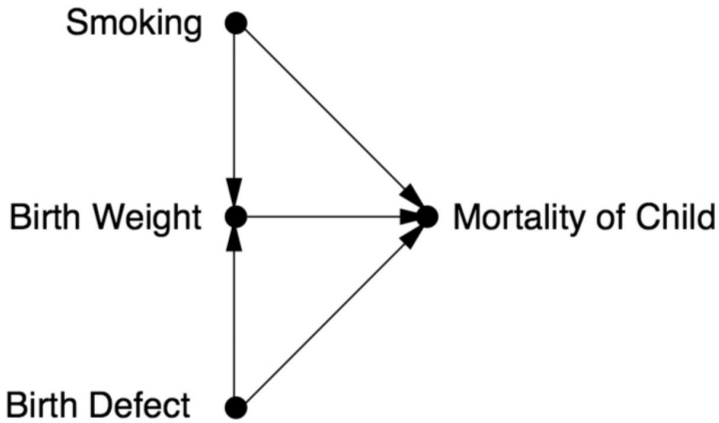
© Théo Jalabert

Explanations

- Other (unmeasured) causes of low birth weight, such as serious or life-threatening genetic abnormalities
- In general there are two possible explanations for low birth weight in one particular baby:
  - it might have a smoking mother, or
  - it might be affected by one of those other (unmeasured) causes.
- If we find out that the mother is a smoker, this explains away the low weight and consequently reduces the likelihood of a serious birth defect.
- But if the mother does not smoke, we have stronger evidence that the cause of the low birth weight is a birth defect, and the baby's prognosis becomes worse

Birth-weight paradox / Causal diagram

© Théo Jalabert



Interpretation of Causal diagram for the Birth-weight paradox

- Birth Weight is a *collider* in above diagram
- Consequently, conditioning on Birth Weight will introduce a spurious correlation through the path Smoking $\rightarrow$ Birth Weight $\leftarrow$ Birth Defect $\rightarrow$ Mortality.
- Solution is either to
  - not condition on Birth Weight, or
  - also condition on Birth Defect (closes backdoor)

Another backdoor example

© Théo Jalabert

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

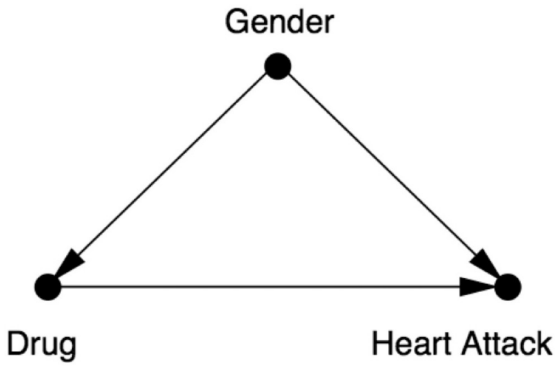
Simpson Paradox

- Observational study about a new drug for reducing the risk of heart attacks
- Risk of heart attack is 21.6% without and 18.3% with new drug
- But for men, new drug increases heart attack probability from 30% to 40%
- So it should be very efficient for women!
- But it also increases risk for women from 5% to 7.5%

$\Rightarrow$ New drug is BBG (bad for men, bad for women, but good for people)

Causal diagram / Simpson paradox

© Théo Jalabert



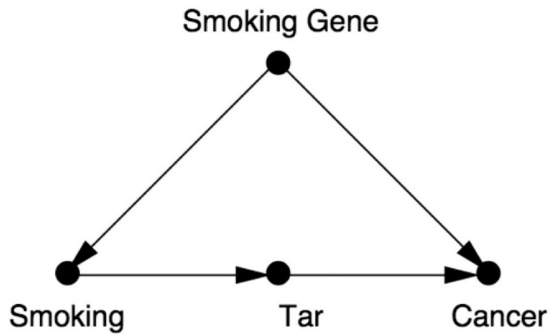
Causal diagram

- Gender has an influence on the risk of heart attacks (men at higher risk)
- In the study, gender also had an influence on the decision whether to take the drug or not (women preferred to take the drug)
- Gender is a confounder, and we have to control for it to block the backdoor path
- Causal effect must be calculated per gender and then averaged (men/women equal in population size)

$\Rightarrow$ New drug is BBB (bad for men, bad for women and bad for people)

Frontdoor paths

© Théo Jalabert



Alternative solution for the smoking problem

- Measure tar deposits in the smokers' lung
- Assume that smoking gene has no effect on tar deposits
- Assume that smoking leads to cancer only through accumulation of tar
- Again Backdoor path cannot be blocked
- But, we can estimate the effect of smoking on tar (Backdoor already blocked by collider cancer)
- Afterwards, we estimate the effect of tar on cancer (again blocked by cancer) and combine both to the total effect

Adjustment formulas

© Théo Jalabert

Backdoor

If Z is a set of covariates, blocking all the backdoor paths for the causal effect $X \to Y$, then

$$\mathbb{P}(Y|do(X)) = \sum_z \mathbb{P}(Z = z)\mathbb{P}(Y|X, Z = z)$$

Frontdoor

If Z is a set of mediators for the causal effect $X \to Y$, then

$$\mathbb{P}(Y|do(X)) = \sum_z \mathbb{P}(Z = z) \sum_x \mathbb{P}(X = x)\mathbb{P}(Y|X = x, Z = z)$$

Frontdoor example

© Théo Jalabert

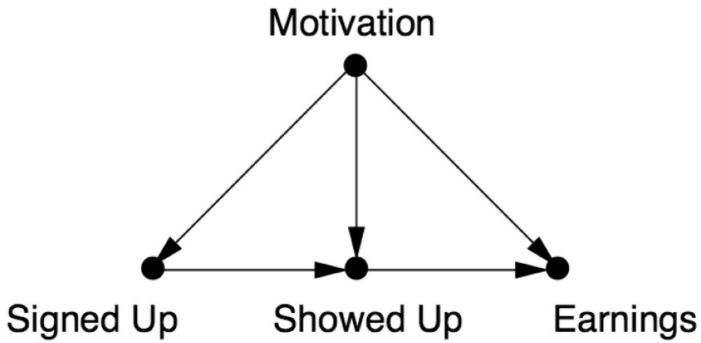
Job Training Partnership Act (JTPA) Study

- Job-training program that provided participants with occupational skills, job-search skills, and work experience.
- Afterwards tracked earnings over the subsequent 18 months
- Study included both, a RCT, where people were randomly assigned to receive services or not, and an observational part, in which people could choose for themselves.

Question: Does job training lead to higher income?

Frontdoor example / Causal diagram

© Theo Jalabert



Causal diagram

- The variable Signed Up records whether a person did or did not register for the program
- The variable Showed Up records whether the enrollee did or did not actually use the services
- No direct arrow from Signed Up to Earnings, because people need to show up to benefit from training
- All kind of confounders summarized in Motivation

Problem: Arrow from Motivation to Showed Up forbids frontdoor criterion

Frontdoor example

© Théo Jalabert

Testing the Frontdoor criterion

- No Effect of Motivation on Showed Up is hard to justify (could be only if it was caused by external events like strike, etc.)
- Glynn and Kashin nevertheless tried the frontdoor approach in their 2014 paper for this scenario
- Know result from RCT, so we can suspect that if middle arrow is weak, frontdoor bias should be small
- Also performed backdoor criterion (controlling for known confounders like Age, Race, Site, ...)
- Result was surprising: Backdoor was wrong by hundreds or thousands of dollars (unobserved confounders!)
- The front-door estimates matched the experimental benchmark almost perfectly

Do-Calculus

© Théo Jalabert

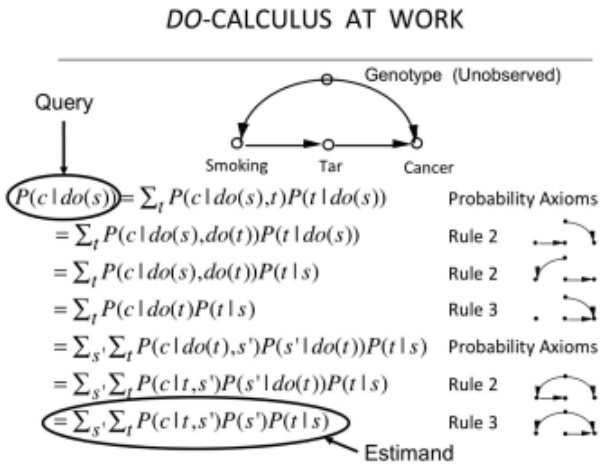
Three fundamental rules of Do-Calculus

Rule 1: If W is independent of Y conditional on Z, then
$\mathbb{P}(Y|do(X), Z, W) = \mathbb{P}(Y|do(X), Z)$

Rule 2: If Z blocks all back-door paths from X to Y, then
$\mathbb{P}(Y|do(X), Z) = \mathbb{P}(Y|X, Z)$

Rule 3: If there are no causal paths from X to Y, then
$\mathbb{P}(Y|do(X)) = \mathbb{P}(Y)$

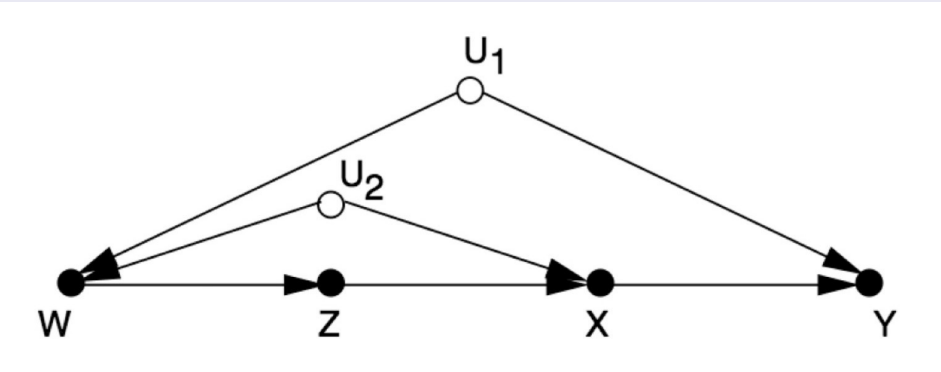
These three rules can prove frontdoor adjustment formula as below, where traditional methods need many pages



Do-Calculus

© Théo Jalabert

Causal diagram, which cannot be solved by back-/frontdoor criterion



Theorem (Pearl and Shpitser, 2006)

A causal effect is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the 3 Do-Calculus rules, that reduces the effect into a standard (i.e. "do"-free)) probability expression involving observed quantities.