

# Bayesian Thinking

Max Kochurov

Moscow State University

August 3, 2022



# Agenda

- ① About Me
- ② Formal stuff
- ③ Motivation
- ④ Bayesian Probability
  - Prior
  - Likelihood
  - Posterior
- ⑤ Model
- ⑥ Discussion
- ⑦ Supplementary
  - Programming



# About Me

- Graduated from
  - BS - MSU EF (2018)
  - MS - Skoltech DS (2020)
- Core developer at PyMC
- Principle Data Scientist at PyMC Labs
- BayesGroup alumni
- Experience with
  - Large scale Deep Learning and Computer Vision
  - Differential geometry for Graph Neural Networks
  - Bayesian Methods for Deep Learning
  - Applied Bayesian Statistics for industry (e.g. AB testing, Bio-Informatics)



# In this course

You'll learn...

- how to think critically about your model
- tools to check the validity of the results
- how to present your results
- non-parametric models for time series

# Grading

The grade consists of

- 60% Homework
- 40% Group project

Grades will be assigned as

- 5 - 85%+
- 4 - 65%+
- 3 - 40%+
- 2 - < 40%

# Why learn Bayesian methods?

- Used in advanced research
  - CBR - Bayesian DSGE ([link](#))
  - Papers using PyMC from google scholar is overwhelming ([link](#))
- Used in top notch industry
  - Marketing at Indigo ([link](#))
  - Drug development at Roche ([link](#))
  - Portfolio Theory at Quantopian ([link](#))
  - Financial Advisory at Everysk ([link](#))
  - Conducting Surveys at Civiqs ([link](#))
- Growth opportunity
  - Links many disciplines and career transitions
  - Hot non-boring job offerings in industry
  - Opens new research possibilities

# Bayesian Probability

$$p(\Theta|D) = \frac{\overbrace{p(D|\Theta)}^{\text{FACT}} \overbrace{p(\Theta)}^{\text{Thinking}}}{p(D)}$$



FACT



$D$  = Data

$\Theta$  = World State

# Prior Distribution



$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta) \overbrace{p(\Theta)}^{\text{Prior}}}{p(\mathcal{D})}$$

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset
- Few to zero relevant prior research
  - no exact match for cannabis case and brain activity
  - additional developing diseases

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset
- Few to zero relevant prior research
  - no exact match for cannabis case and brain activity
  - additional developing diseases
- Scarce, hard to obtain data

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset
- Few to zero relevant prior research
  - no exact match for cannabis case and brain activity
  - additional developing diseases
- Scarce, hard to obtain data
- Classical econometrics fails (16 data points in the group)

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset
- Few to zero relevant prior research
  - no exact match for cannabis case and brain activity
  - additional developing diseases
- Scarce, hard to obtain data
- Classical econometrics fails (16 data points in the group)
- Expert knowledge feels important

# Case Study: Where do priors come from?

Authors: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Bayesian Econometrics example

## Policy question

Should we increase cannabis control for adolescents?

- Drugs long term influence on brain activity after early onset
- Few to zero relevant prior research
  - no exact match for cannabis case and brain activity
  - additional developing diseases
- Scarce, hard to obtain data
- Classical econometrics fails (16 data points in the group)
- Expert knowledge feels important
- Statistics is required for an informed decision

# Quick intro

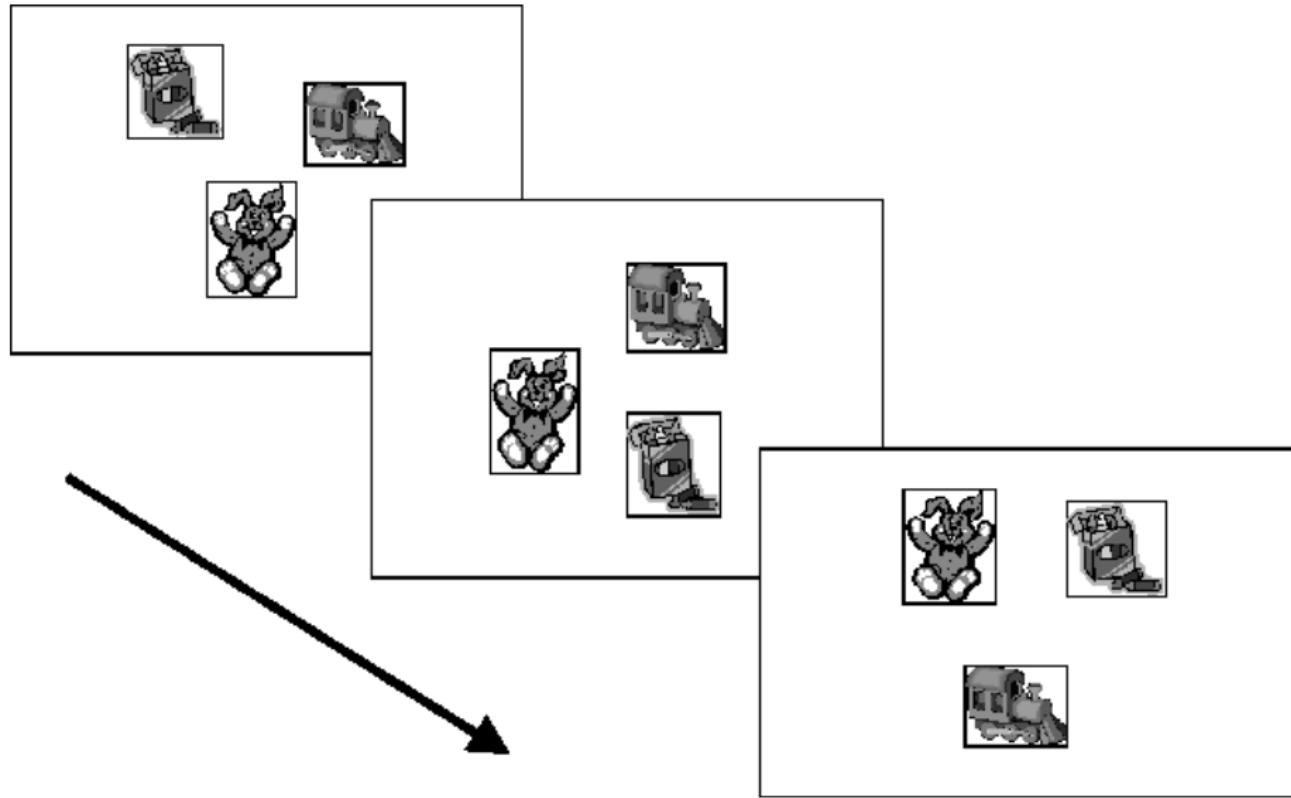
Measuring the existence and severity of cannabis usage

- You can measure brain development with a Game ( Self Ordered Pointing Test [1])
- Cannabis use is checked for participants
- Adolescents pass the Game 2 times a year
- The results are compared between the **heavy** and **light** cannabis users

Leftover questions

- How does the amount of cannabis used affects the brain development?
- What age is sufficient to minimize the effect of usage?
- What policy should be used to minimize the effect?

# The Game



# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

① Knowledge before seeing any data



② Prior research results



③ Expert knowledge



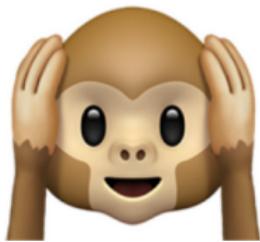
④ Constraints

⑤ Model properties

# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

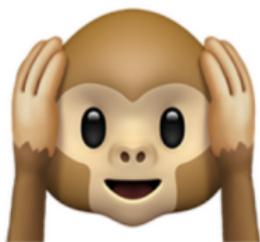
- ① Knowledge before seeing any data
  - Practical range for SOPT measure and growth rates
- ② Prior research results
- ③ Expert knowledge
- ④ Constraints
- ⑤ Model properties



# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

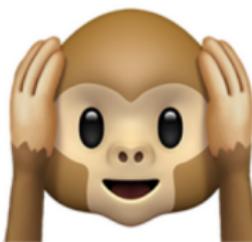
- ① Knowledge before seeing any data
  - Practical range for SOPT measure and growth rates
- ② Prior research results
  - effect size of cannabis use was mixed with other diseases or had missing information
  - "13 out of 693 articles yielded useful information"
- ③ Expert knowledge
- ④ Constraints
- ⑤ Model properties



# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

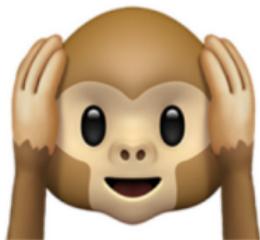
- ① Knowledge before seeing any data
  - Practical range for SOPT measure and growth rates
- ② Prior research results
  - effect size of cannabis use was mixed with other diseases or had missing information
  - "13 out of 693 articles yielded useful information"
- ③ Expert knowledge
  - Relation of diseases and behaviour to cannabis use
  - Prior research was reviewed by experts
- ④ Constraints
- ⑤ Model properties



# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

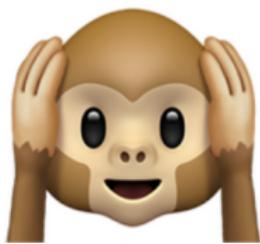
- ① Knowledge before seeing any data
  - Practical range for SOPT measure and growth rates
- ② Prior research results
  - effect size of cannabis use was mixed with other diseases or had missing information
  - "13 out of 693 articles yielded useful information"
- ③ Expert knowledge
  - Relation of diseases and behaviour to cannabis use
  - Prior research was reviewed by experts
- ④ Constraints
  - Slope (SOPT development rate) is more positive than negative
- ⑤ Model properties



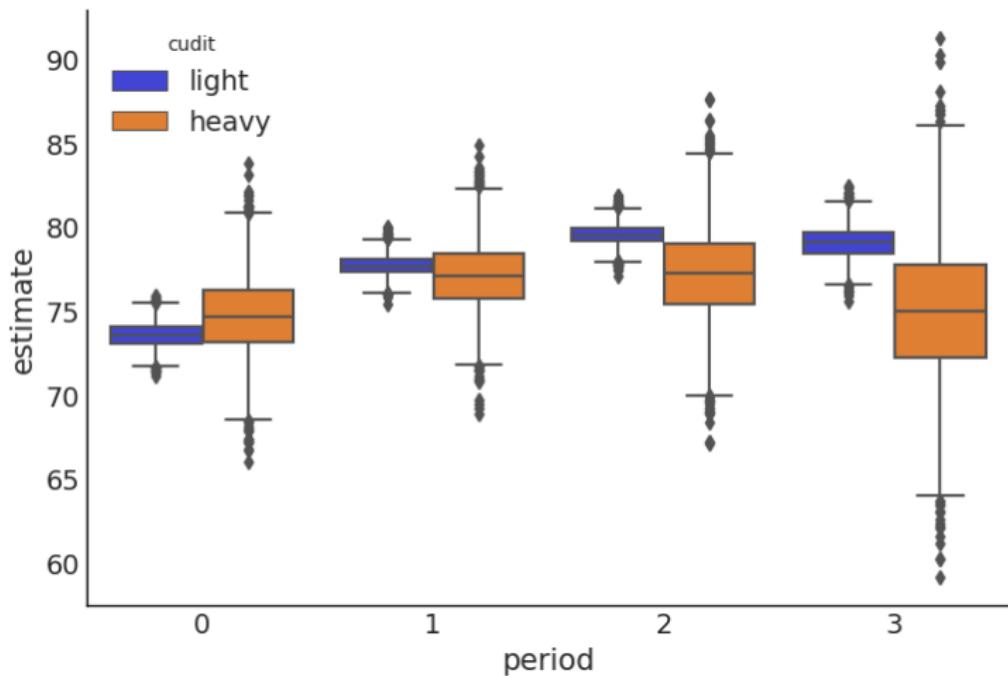
# Case Study: Prior Distribution

To develop a prior researchers combined many sources of information

- ① Knowledge before seeing any data
  - Practical range for SOPT measure and growth rates
- ② Prior research results
  - effect size of cannabis use was mixed with other diseases or had missing information
  - "13 out of 693 articles yielded useful information"
- ③ Expert knowledge
  - Relation of diseases and behaviour to cannabis use
  - Prior research was reviewed by experts
- ④ Constraints
  - Slope (SOPT development rate) is more positive than negative
- ⑤ Model properties
  - Sign of the quadratic term



# Results



We'll reproduce this plot on the seminar

# Common Issues

In the research the prior was defended

- ① Prior is subjective
- ② Prior specification is unclear
- ③ Prior is incorrectly specified

# Common Issues

In the research the prior was defended

- ① Prior is subjective
  - Informed prior was compared to uninformed one
- ② Prior specification is unclear
- ③ Prior is incorrectly specified

# Common Issues

In the research the prior was defended

- ① Prior is subjective
  - Informed prior was compared to uninformed one
- ② Prior specification is unclear
  - The log book was provided with a full description of the choice
  - <https://osf.io/aw8fy/>
- ③ Prior is incorrectly specified

# Common Issues

In the research the prior was defended

- ① Prior is subjective
  - Informed prior was compared to uninformed one
- ② Prior specification is unclear
  - The log book was provided with a full description of the choice
  - <https://osf.io/aw8fy/>
- ③ Prior is incorrectly specified
  - There are still some issues with the analysis, we'll review them on the seminar

# Likelihood Distribution

**FACT**

$$p(\Theta | \mathcal{D}) = \frac{\overbrace{p(\mathcal{D} | \Theta)} p(\Theta)}{p(\mathcal{D})}$$

# Case Study: AB Test

You sell nuts. You want sell more nuts! How to increase sales?

- increase purchase probability



- increase order size



# Case Study: AB Test

You sell nuts. You want sell more nuts! How to increase sales?

- increase purchase probability
  - Create a banner about healthy food
  - Add a banner with recipes
  - Improve the layout
- increase order size



# Case Study: AB Test

You sell nuts. You want sell more nuts! How to increase sales?

- increase purchase probability
  - Create a banner about healthy food
  - Add a banner with recipes
  - Improve the layout
  
- increase order size
  - Lower the price
  - Increase quality
  - Make better packaging



# Case Study: AB Test

You sell nuts. You want sell more nuts! How to increase sales?

- increase purchase probability
  - Create a banner about healthy food
  - Add a banner with recipes
  - Improve the layout
  
- increase order size
  - Lower the price
  - Increase quality
  - Make better packaging

What is better?

AB testing can answer the question



# Not all customers buy nuts



- A significant portion of data are just zeros
- A classical 2 sample t-test assumes normality, not our case
- Researchers admit t-test weaknesses in these cases [2]

# Not all customers buy nuts



- A significant portion of data are just zeros
- A classical 2 sample t-test assumes normality, not our case
- Researchers admit t-test weaknesses in these cases [2]

## Solution

Think of a non-normal likelihood

# Zero inflation

## Zero Inflation

Data property, when a significant portion of data is exactly zero

1	0	1	0	0	0	0	1	1	1
0	1	0	0	1	0	0	1	1	0
0	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	1
0	1	1	0	1	1	0	0	1	0
1	0	1	0	0	1	0	1	0	1
1	0	1	1	1	0	1	1	1	1
0	0	0	0	1	0	0	1	1	1
0	0	0	1	0	0	1	1	0	0
0	1	0	0	1	1	0	1	0	1

Examples:

- Wait times in a queue (no queue is zero)
- Defects on a production line (no defects is zero)
- Rain water level (sunny weather is 0 water level)
- Purchase order statement (no nuts is zero)

# Zero Inflated Distribution

## Wisdom

Any distribution can be made zero inflated

**Example:** Zero Inflated Gamma.

We'll have parameters  $\alpha, \beta$  (from Gamma) and  $p$  - probability of non-zero

**Sampling:**

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$

# Zero Inflated Distribution

## Wisdom

Any distribution can be made zero inflated

**Example:** Zero Inflated Gamma.

We'll have parameters  $\alpha, \beta$  (from Gamma) and  $p$  - probability of non-zero

**Sampling:**

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$

**Log Probability Density Function**

$$\log p(x | p, \alpha, \beta) = \begin{cases} \log(1 - p) & , x = 0 \\ \log(p) \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} & , x > 0 \end{cases}$$

# Zero Inflated Distribution

## Wisdom

Any distribution can be made zero inflated

**Example:** Zero Inflated Gamma.

We'll have parameters  $\alpha, \beta$  (from Gamma) and  $p$  - probability of non-zero

**Sampling:**

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$

**Log Probability Density Function**

$$\log p(x | p, \alpha, \beta) = \begin{cases} \log(1 - p) & , x = 0 \\ \log(p) \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} & , x > 0 \end{cases}$$

# Zero inflation as a mixture

## Wisdom

Zero inflation is a special case of a Mixture

Components:

- ① Constant(0)
- ② Gamma( $\alpha, \beta$ )

Mixture probability is  $p$

$$\text{ZI-Gamma}(p, \alpha, \beta) \equiv \text{Mixture}([1 - p, p], [\text{Constant}(0), \text{Gamma}(\alpha, \beta)])$$

# Back to the example

make an order  $\sim \text{Bernoulli}(p)$

order amount  $\sim \begin{cases} \text{Gamma}(\alpha, \beta) & , \text{make an order} = 1 \\ 0 & , \text{make an order} = 0 \end{cases}$



# Takeouts

- Good likelihood helps to get better sense of the problem
  - split purchase probability and purchase amount
  - more possibilities over a classical t-test
- Understanding a problem is a first step to a good likelihood

# Posterior Distribution

$$\underbrace{p(\Theta | \mathcal{D})}_{\text{Posterior}} = \frac{p(\mathcal{D} | \Theta) p(\Theta)}{p(\mathcal{D})}$$

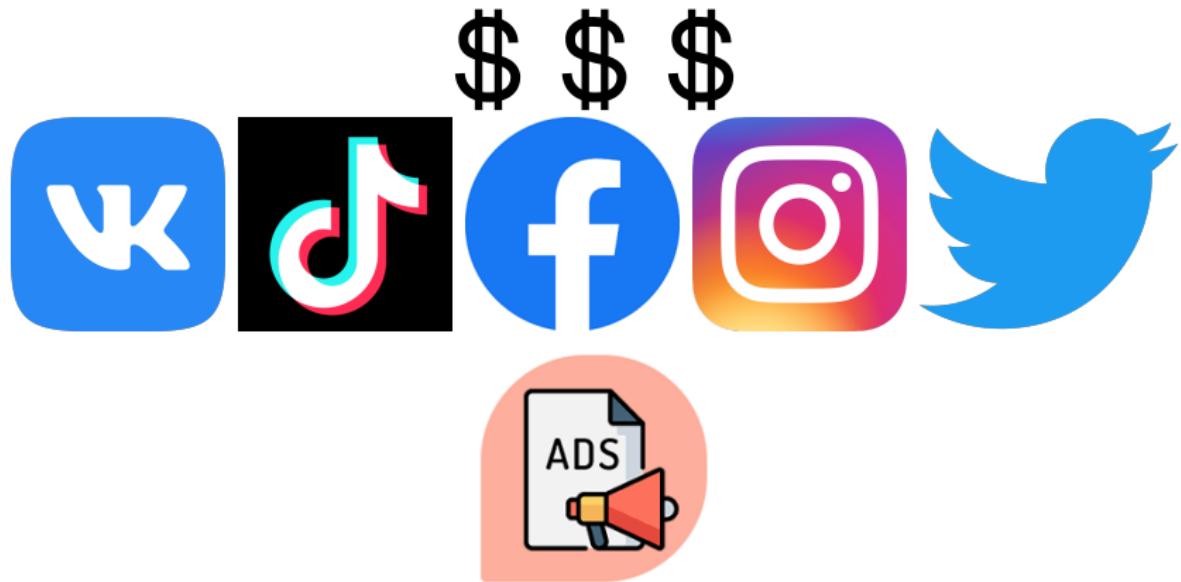


# Posterior Distribution

$$\begin{aligned} p(\text{what you think} | \text{data}) \\ \propto p(\text{data} | \text{what you think}) p(\text{what you think}) \end{aligned}$$



# Case Study: Marketing



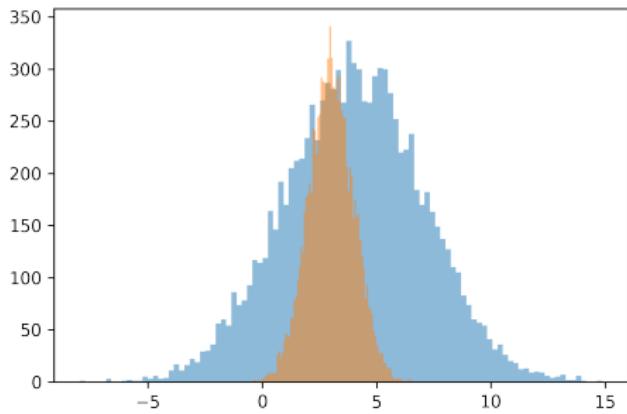
## Marketing Mix Model

MMM - helps to evaluate marketing channels from historical data

# Uncertainty

Marketing Mix Models ([read more here](#))

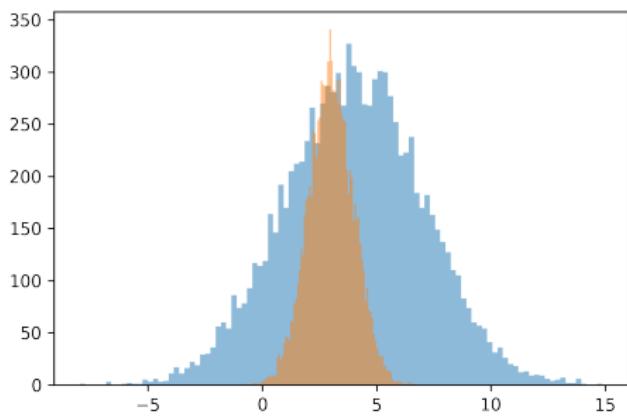
- How valuable is additional \$1000 invested in VK? Or Facebook?
- How certain is the model estimation?
- High value, high uncertainty or low value low uncertainty?
- How to allocate money?



# Uncertainty

Marketing Mix Models ([read more here](#))

- How valuable is additional \$1000 invested in VK? Or Facebook?
- How certain is the model estimation?
- High value, high uncertainty or low value low uncertainty?
- How to allocate money?



## Takeout

Uncertainty helps to make more informed decision

# Recap

In Bayesian framework you have:

- Prior = What I think the problem is
- Likelihood = What the facts I have
- Posterior = What the problem actually is *given priors and data*

# The Model

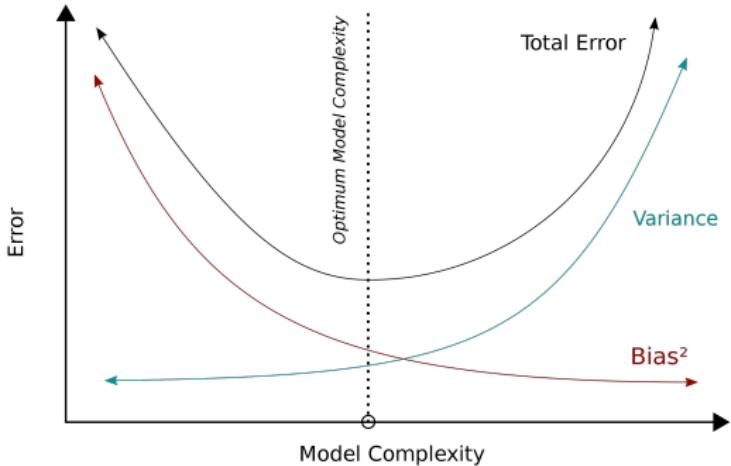
$$p_{\mathcal{M}}(\Theta | \mathcal{D}) = \frac{p_{\mathcal{M}}(\mathcal{D} | \Theta) p_{\mathcal{M}}(\Theta)}{p_{\mathcal{M}}(\mathcal{D})}$$

Treat the model as a "one of many" descriptions of the problem.

# Bias - Variance trade-off

Getting to a good model

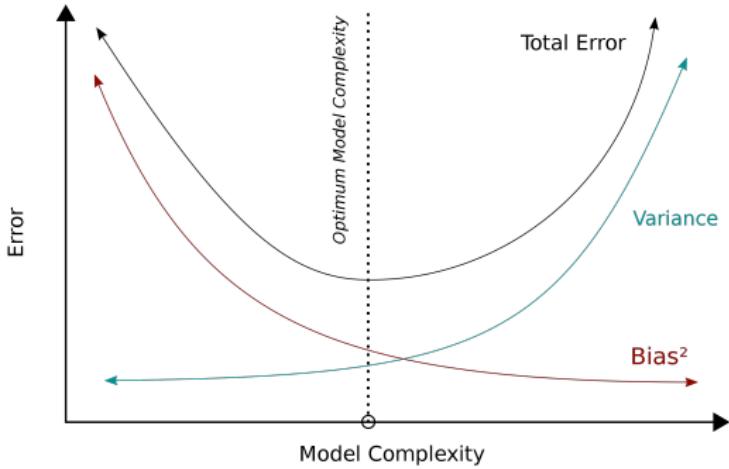
- ① Start with a over-simplified model
- ② Make sure it samples well
- ③ Increase the complexity
- ④ ...
- ⑤ Choose the best model using cross validation



# Bias - Variance trade-off

Getting to a good model

- ① Start with a over-simplified model
- ② Make sure it samples well
- ③ Increase the complexity
- ④ ...
- ⑤ Choose the best model using cross validation



## Common mistake

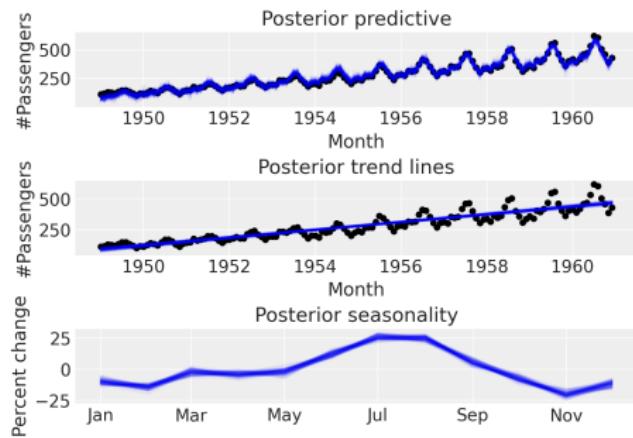
Starting with a complicated model make debugging painful

# Case Study: generalized additive models

$$y(t) = \underbrace{g(t)}_{\text{trend}} + \underbrace{s(t)}_{\text{seasonality}} + \underbrace{r(X_t)}_{\text{regressors}} + \varepsilon_t$$

Adding more and more complexity

- ① start with a simple trend model
- ② add seasonality
- ③ add fine seasonality details, holidays or other features



# When Bayes?

Bayesian modeling does not fit all the use cases at once

- It requires extra skills compared to classical machine learning
- You might not need the extra "uncertainty"

# When Bayes?

Bayesian modeling does not fit all the use cases at once

- It requires extra skills compared to classical machine learning
- You might not need the extra "uncertainty"

## Proposition

Bayesian modeling starts when fit-predict is useless

# When Bayes?

Bayesian modeling does not fit all the use cases at once

- It requires extra skills compared to classical machine learning
- You might not need the extra "uncertainty"

## Proposition

Bayesian modeling starts when fit-predict is useless

- You have interpretable confidence intervals
- You have flexibility and control over the model
- You have reliable low data applications
- You ultimately HAVE to understand the model

# PyMC

- ① Pure Python!
- ② Automated inference
- ③ No complicated formulas for MCMC!
- ④ Visualizations with ArviZ
- ⑤ Reproducible research
- ⑥ Used in industry applications
- ⑦ Huge community
- ⑧ Active development

<https://github.com/pymc-devs/pymc>



# Discussion Time

- Bayesian model is rarely a default choice
- Frequentist methods are good for fit-predict
- Bayesian modelling is a time consuming work
- Some problems are easier to solve without Bayesian methods

# References I

-  L. Cragg and K. Nation.  
Self-ordered pointing as a test of working memory in typically developing children.  
*Memory (Hove, England)*, 15:526–35, 08 2007.
-  F. McElduff, M. Cortina-Borja, S.-K. Chan, and A. Wade.  
When t-tests or wilcoxon-mann-whitney tests won't do.  
*Advances in Physiology Education*, 34(3):128–133, 2010.  
PMID: 20826766.
-  M. Zondervan-Zwijnenburg, M. Peeters, S. Depaoli, and R. V. de Schoot.  
Where do priors come from? applying guidelines to construct informative priors in small sample research.  
*Research in Human Development*, 14(4):305–320, 2017.