Bayesian Inference in action

Deep-dive into Facebook Prophet

Jovan Krajevski

22 May 2025

Table of Contents

1 Introduction

Frequentist vs Bayesian inference

- Objective vs Subjective
- Frequency vs Belief
- Point estimate vs Posterior distribution
- Confidence intervals vs Credible intervals

When to use Bayesian inference?

- Prior knowledge available
- Point estimates will not suffice
- ***Hierarchical modeling***

Table of Contents

2 The mandatory coin toss problem

The problem

2 The mandatory coin toss problem

- X_i - 1 (heads) or 0 (tails)
- $\{X_1, \ldots, X_n\}$ - a sample of n coin tosses
- Is the coin balanced?

Table of Contents

2 The mandatory coin toss problem

The frequentist model

2 The mandatory coin toss problem

- p - the probability of getting 1
  - toss the coin an infinite number of times
  - the ***frequency*** of 1s is p
- $X_i \sim Bernoulli(p)$

$$P(X_i = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- $Y = \sum_{i=1}^{n} X_i \implies Y \sim B(n, p)$

$$P(Y_i = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Estimating p from a sample

- $\widehat{p}$ - function of $\{X_1, \ldots, X_n\}$ called an ***estimator***
- $\mathcal{L}(p) = P(\{X_1, \ldots, X_n\}|p)$ - the ***likelihood*** function
- $\widehat{p} = \arg\max_p \mathcal{L}(p)$ - the ***maximum likelihood estimator*** (MLE)
- For the coin toss problem, the MLE is:

$$\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

What is a confidence interval?

- MLE is a ***point estimate***
  - We know nothing about other possible values of p
  - We have no notion of confidence (belief, certainty)...
  - ...because p ***is not*** a random variable
- What about the $c\%$ ***confidence interval***?
  - $\{\{X_{1,1}, X_{1,2}, ..., X_{1,n_1}\}, \{X_{2,1}, X_{2,2}, ..., X_{2,n_2}\}, ...\}$
  - $c\%$ confidence interval for each sample
  - $c\%$ of the "$c\%$ confidence intervals" contain the true value of p
  - Confidence in our method, not in our estimate!

Table of Contents

2 The mandatory coin toss problem

Everything is random!

2 The mandatory coin toss problem

- Why is p not a random variable?
- Random variables have probability distribution
- Probability distribution give us ***real*** confidence intervals - ***credible intervals***
- p is not really random - our belief in p is random!

Life before the sample... and after too!

- If p is a random variable:
  - $P(p = x)$ - probability distribution before the sample - ***prior***
  - $P(p = x|\{X_1, \ldots, X_n\})$ - probability distribution after the sample - ***posterior***
- How do we calculate the posterior, given the sample and the prior?
  - $P(\{X_1, \ldots, X_n\}|p)$ - the likelihood w.r.t. the prior

The Bayes theorem

$$
\begin{aligned}
P(p = x|\{X_1, \dots, X_n\}) &= \frac{P(p = x, \{X_1, \dots, X_n\})}{P(\{X_1, \dots, X_n\})} \\
&= \frac{P(p = x)P(\{X_1, \dots, X_n\}|p = x)}{P(\{X_1, \dots, X_n\})} \\
\textbf{posterior} &= \frac{\textbf{prior} \times \textbf{likelihood}}{\textbf{evidence}}
\end{aligned}
$$

- $P(\{X_1, \dots, X_n\})$ is difficult to calculate, but it is constant w.r.t. the sample

$$\textbf{posterior} \propto \textbf{prior} \times \textbf{likelihood}$$

Now back to tossing coins...

2 The mandatory coin toss problem

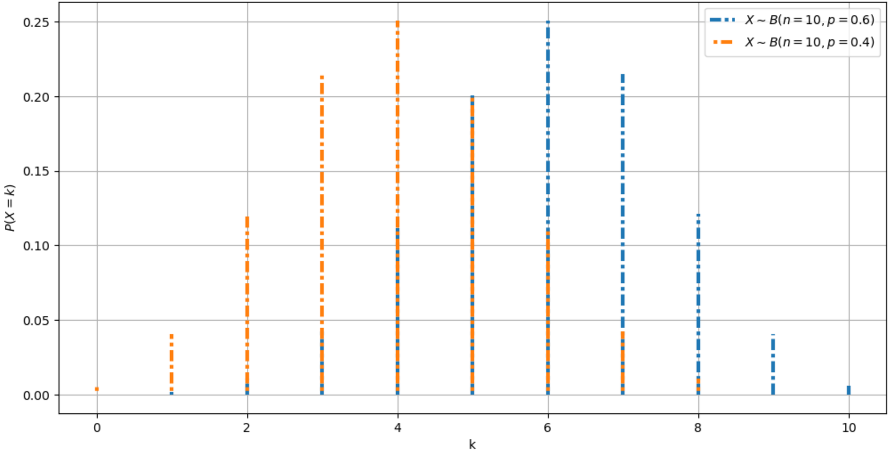
- Let's model $Y = \sum_{i=1}^{n} X_i$
- We already know the likelihood function:

$$P(Y = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- But how do we choose a prior?

The binomial likelihood

2 The mandatory coin toss problem



A different perspective

- Three "values" in the likelihood function
  - Y - number of 1s
  - n - number of trials
  - p - probability of getting 1
- Y is a random variable, n and p are fixed

$$P(Y = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- What if p was not fixed (a random variable)?

$$P(p = x; n, k) \propto x^k (1-x)^{n-k}$$

Introducing the Beta prior

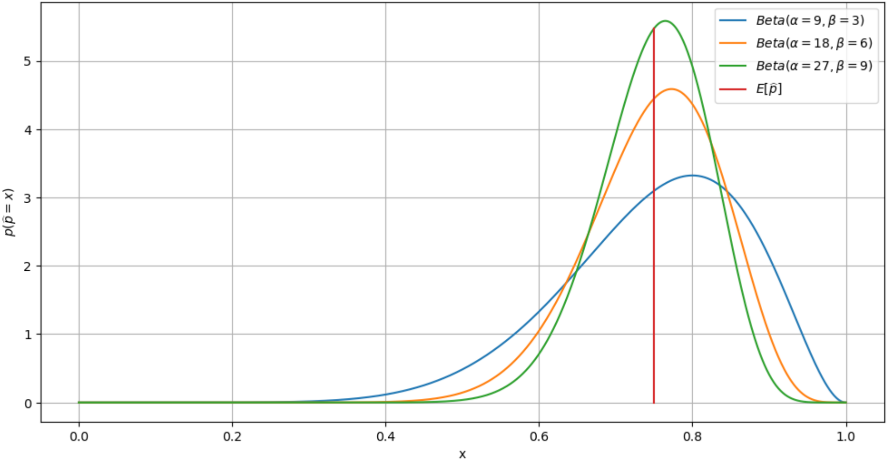
$$P(p = x; n, k) \propto x^k (1-x)^{n-k}$$

- How to replace $\propto$ with $=$? Scale so that the probability density integrates to 1!

$$P(p = x; n, k) = \frac{x^k (1-x)^{n-k}}{B(k+1, n-k+1)}$$

$$\alpha = k+1, \beta = n-k+1 \implies P(p = x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

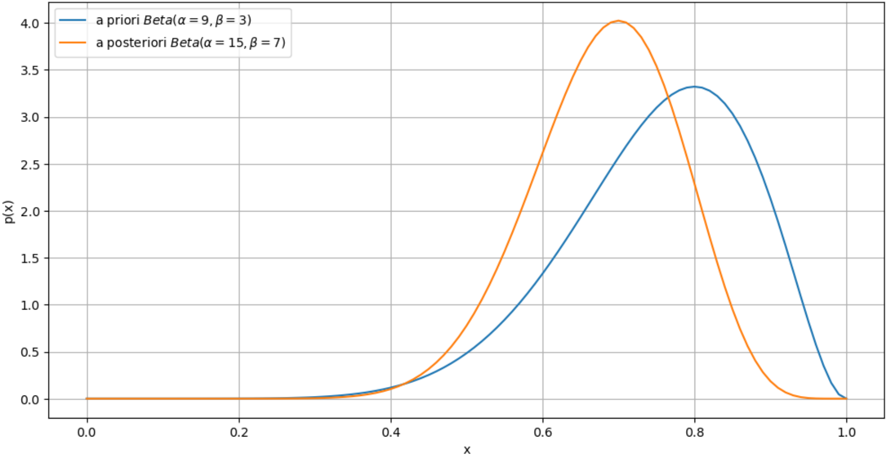
The Beta prior



Calculating the posterior

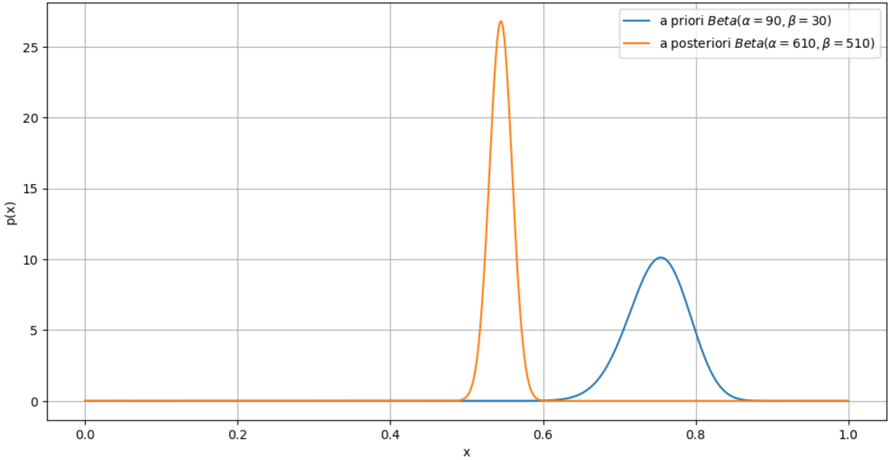
- ***Conjugate priors***
- Prior is $Beta(\alpha, \beta)$
- Likelihood is $B(n, p)$
- Sample is $Y = k$, n trials
- Posterior is $Beta(\alpha + k, \beta + n - k)$

Prior → Posterior



Prior → Posterior

2 The mandatory coin toss problem



Prior → Posterior

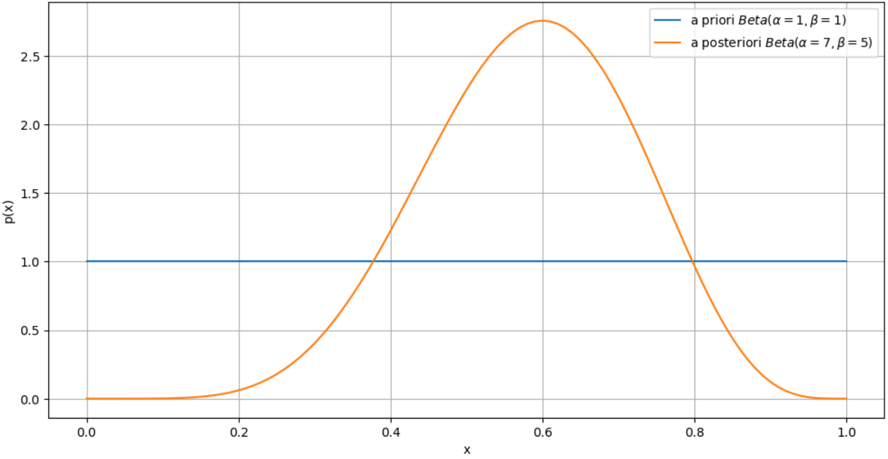


Table of Contents

3 Bayesian inference techniques

Maximum a posteriori (MAP) estimate

- Point estimate
  - mode of the posterior distribution
- Regularized generalization of MLE
  - L1 regularization $\Leftrightarrow$ Laplace prior + MAP
  - L2 regularization $\Leftrightarrow$ Gaussian prior + MAP
- Fast, usually optimized with a gradient-based approach
  - L-BFGS-B

Markov chain Monte Carlo (MCMC) sampling

- Explore full posterior
  - Markov assumption: the future is conditionally independent of the past, given the present
  - Monte Carlo: generate draws from the posterior
- Slow, hard to run in parallel
  - we must draw at step t before can draw at step $t + 1$

MCMC algorithms

- Metropolis-Hastings (MH)
  - original, slow convergence
- NUTS
  - gradients used to detect high density regions

MCMC template

1. Calculate posterior of current point $P(p = x_{now}|Y)$
2. Calculate posterior of next point $P(p = x_{next}|Y)$
3. If $\frac{P(p=x_{next}|Y)}{P(p=x_{now}|Y)} > u; u \sim Uniform(0, 1)$
   3.1 Add next point to trace
   3.2 Set current point to next point
4. Repeat until convergence

Variational Inference (VI)

- Middle-ground approach
- Analytical approximation to the true posterior
  - does not capture all essential characteristics
- Steps:
  - propose variational distribution
  - minimize KL divergence between true posterior and variational distribution
  - draw from variational distribution

VI algorithms

- ADVI
  - factorized Gaussian approximation
- FullRank ADVI:
  - ADVI + parameter correlations through a full-rank Gaussian

MAP vs MCMC vs VI

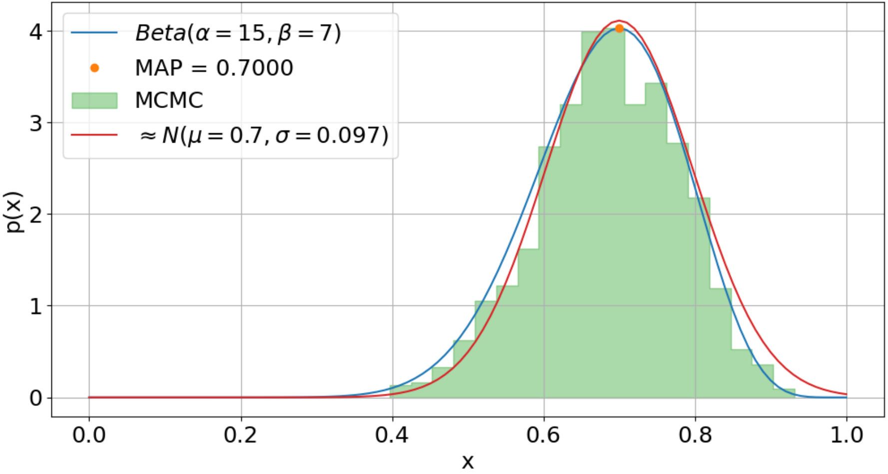


Table of Contents

4 Let's talk about priors

Types of priors

- Informative priors
  - Derived from expert knowledge
  - No formal method to get to them
- Generic weakly informative priors
- Weakly informative priors
  - Principled regularization
- Vague priors
- Flat priors
  - Frequentist approach + posterior exploration

So what is the Normal distribution?

$$P(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $(x - \mu)^2$ - we are modeling the square distance from what we think our parameter should be!
- $2\sigma^2$ - controls the spread of the bell shaped curve!
- $\frac{1}{\sqrt{2\pi\sigma^2}}$ - a scaling monstrocity to make this integrate to 1!

What happens when we use a Normal prior?

$$
\begin{aligned}
\textbf{posterior} &\propto \textbf{prior} \times \textbf{likelihood} \\
\textbf{posterior} &\propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \times \textbf{likelihood} & / \log \\
\textbf{logposterior} &\propto -\frac{1}{2\sigma^2}(x-\mu)^2 + \textbf{loglikelihood}
\end{aligned}
$$

- The scaling factor does not matter!
- If $\mu = 0$, this is L2 regularization with $\lambda = \frac{1}{2\sigma^2}$

$$\mathcal{L}_{L2} = \mathcal{L} + \lambda \|x\|_2^2$$

So what is the Laplace distribution?

$$P(X = x; \mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

- $|x - \mu|$ - we are modeling the absolute distance from what we think our parameter should be!
- b - controls the spread of the bell shaped curve!
- $\frac{1}{2b}$ - scaling to make this integrate to 1!

What happens when we use a Laplace prior?

$$\begin{aligned}
\textbf{posterior} &\propto \textbf{prior} \times \textbf{likelihood} \\
\textbf{posterior} &\propto \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \times \textbf{likelihood} && / \log \\
\textbf{logposterior} &\propto -\frac{1}{b}|x-\mu| + \textbf{loglikelihood}
\end{aligned}$$

- If $\mu = 0$, this is L1 regularization with $\lambda = \frac{1}{b}$

$$\mathcal{L}_{L1} = \mathcal{L} + \lambda \|x\|_1$$

Table of Contents

5 Facebook Prophet

What is a time series?

- Given a set of time-points $T = \{t^{(k)} : k \in \mathbb{N}\} \subseteq \mathbb{R}$
- And a function $y : T \to \mathbb{R}$
- A time series is the set:

$$X = \{y(t) : t \in T\}$$

The Facebook Prophet model

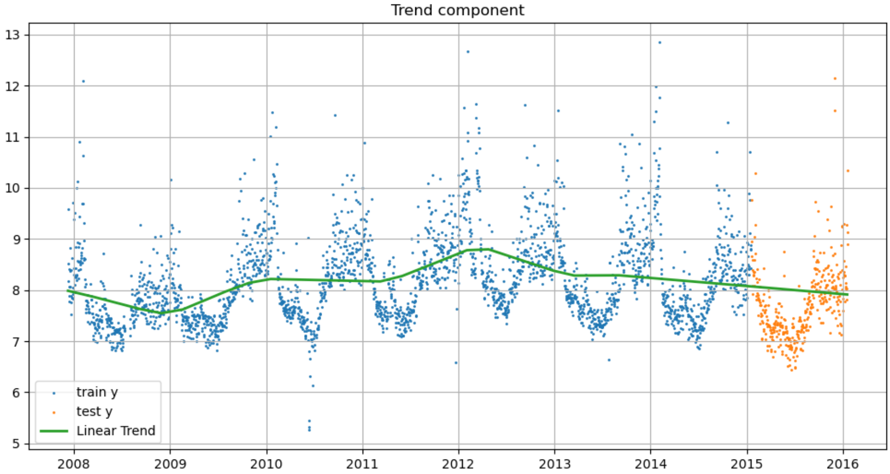
- Generalized additive model:

$$y(t;\boldsymbol{\gamma},\boldsymbol{\theta}) = g(t;\boldsymbol{\gamma}) + \sum_{j=1}^{m} s(p^{(j)}, t; \boldsymbol{\theta}^{(j)}) + \epsilon^{(t)}$$

- g - trend parametrized with $\boldsymbol{\gamma}$
- s - seasonality parametrized with $\boldsymbol{\theta}^{(j)}$ for period $p^{(j)}$

$$P = \{p^{(j)} : j \in \mathbb{N}, 1 \leq j \leq m\} \subseteq \mathbb{R}$$
$$\forall k, j \in \mathbb{N}, j \leq m$$
$$s(p^{(j)}, t_k) = s(p^{(j)}, t_k - p^{(j)})$$

Visual representation of Facebook Prophet



Visual representation of Facebook Prophet

5 Facebook Prophet

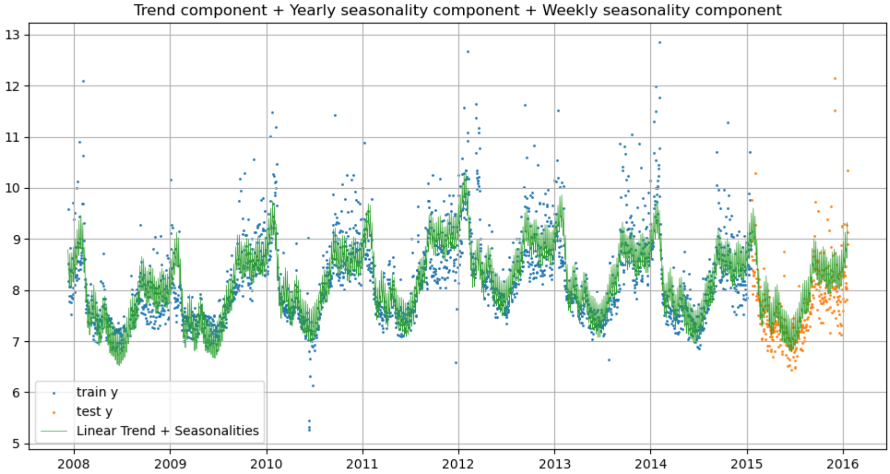
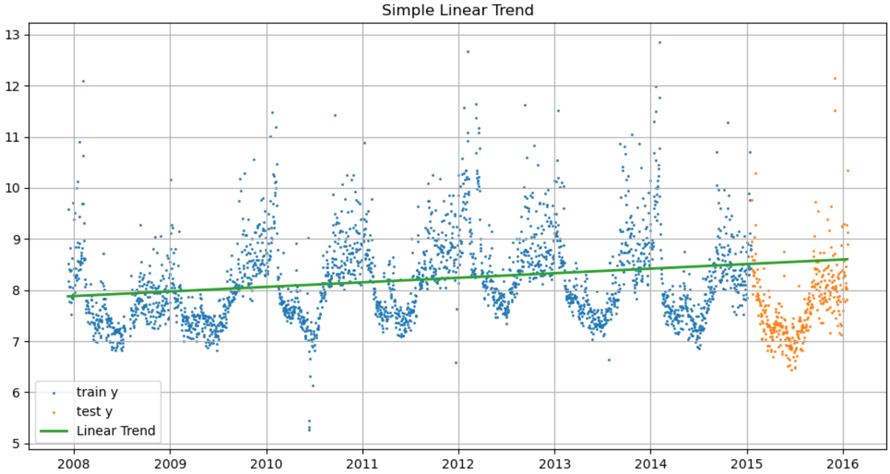


Table of Contents

5 Facebook Prophet

The simple form - Linear Trend

5 Facebook Prophet

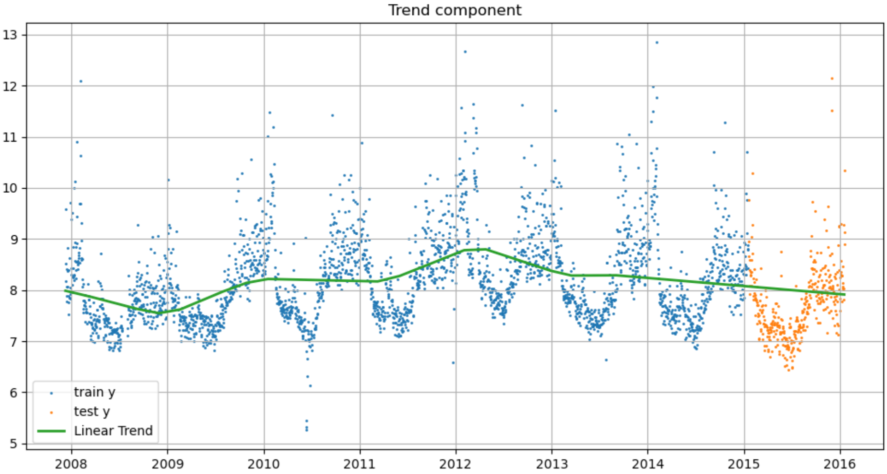


The simple form - Linear Trend

- Two parameters: slope and intercept

$$
\begin{aligned}
w &\sim \mathcal{N}(0, 5) \\
b &\sim \mathcal{N}(0, 5) \\
\boldsymbol{\gamma} &= \begin{pmatrix} w & b \end{pmatrix} \\
g(t; \boldsymbol{\gamma}) &= wt + b
\end{aligned}
$$

Introducing change points



Introducing change points

- Points in which w can chage
- $S \in \mathbb{N}$ - number of change points
- $\mathbf{s} \in T^S$ - column vector of change points
  - Set at equal distance
- $\boldsymbol{\delta} \in \mathbb{R}^S$ - change in w at points $\mathbf{s}$
  - Random vector
  - $\boldsymbol{\delta} \sim Laplace(0, 0.05)$

The slope at every time-point

- $\mathbf{t} = \begin{pmatrix} t^{(1)} & t^{(2)} & \dots t^{(|T|)} \end{pmatrix}$ - row vector representation of T
- $\mathbf{w} \in \mathbb{R}^{|T|}$ - column vector representation of the slope at every time-point

$$w_k = w + \sum_{l=1}^{t^{(k)} > s_l} \delta_l$$

Vectorizing the slope calculation

- $\mathbf{a}_k \in \{0, 1\}^S$ - row vector

$$a_k^{(l)} = \begin{cases} 1 & t^{(k)} > s_l \\ 0 & t^{(k)} \leq s_l \end{cases} \implies w_k = w + \mathbf{a}_k \boldsymbol{\delta}$$

- $\mathbf{A} \in \mathbb{R}^{|T| \times S}$ - matrix, each row is $\mathbf{a}_k$

$$\mathbf{w} = w + \mathbf{A} \boldsymbol{\delta}$$

What about the intercept?

- $\mathbf{b} \in \mathbb{R}^{|T|}$ - column vector representation of the intercept at every time-point

$$g(t;\boldsymbol{\gamma}) = \begin{cases} wt + b & t < s_1 \\ (w + \delta_1)t + b_1 & t \in [s_1, s_2) \\ \dots & \\ (w + \sum_{l=1}^{m} \delta_l)t + b_m & t \in [s_m, s_{m+1}) \\ \dots & \\ (w + \sum_{l=1}^{S} \delta_l)t + b_S & t \geq s_S \end{cases}$$

Find b!

- The trend is continuous if:

$$(w + \sum_{l=1}^{m-1} \delta_l) s_m + b_{m-1} = (w + \sum_{l=1}^{m} \delta_l) s_m + b_m$$

- Solve for b_m:

$$b_m = b_{m-1} - \delta_m s_m$$

- Special case:

$$w s_1 + b = (w + \delta_1) s_1 + b_1 \implies b_1 = b - \delta_1 s_1$$

Vectorize the intercept calculation

- Recursively apply

$$b_1 = b - \delta_1 s_1; \qquad b_m = b_{m-1} - \delta_m s_m$$

$$\implies b_m = b - \sum_{l=1}^{m} \delta_l s_l = b - \sum_{l=1}^{t^{(k)} > s_l} \delta_l s_l$$

- Now we can use $\mathbf{A}$

$$b_m = b - \mathbf{a}_m(\boldsymbol{\delta} \odot \mathbf{s}) \implies \mathbf{b} = b - \mathbf{A}(\boldsymbol{\delta} \odot \mathbf{s})$$

Final piece-wise linear trend form

- All parameters

$$\boldsymbol{\gamma} = \begin{pmatrix} w & b & \delta_1 & \dots & \delta_S \end{pmatrix}$$

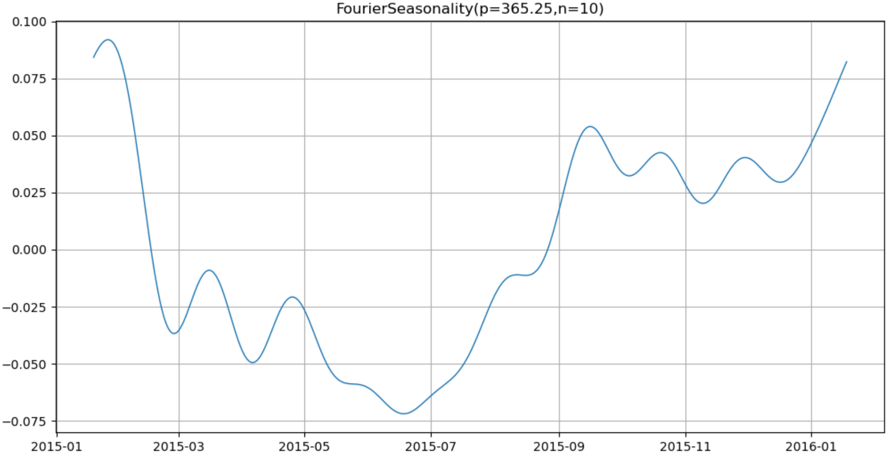
- $\mathbf{g} \in \mathbb{R}^{|T|}$ - column vector representation of the linear trend at every time-point

$$\begin{aligned} \mathbf{g} &= \mathbf{w} \odot \mathbf{t}^T + \mathbf{b} \\ &= (w + \mathbf{A}\boldsymbol{\delta}) \odot \mathbf{t}^T + b - \mathbf{A}(\boldsymbol{\delta} \odot \mathbf{s}) \end{aligned}$$

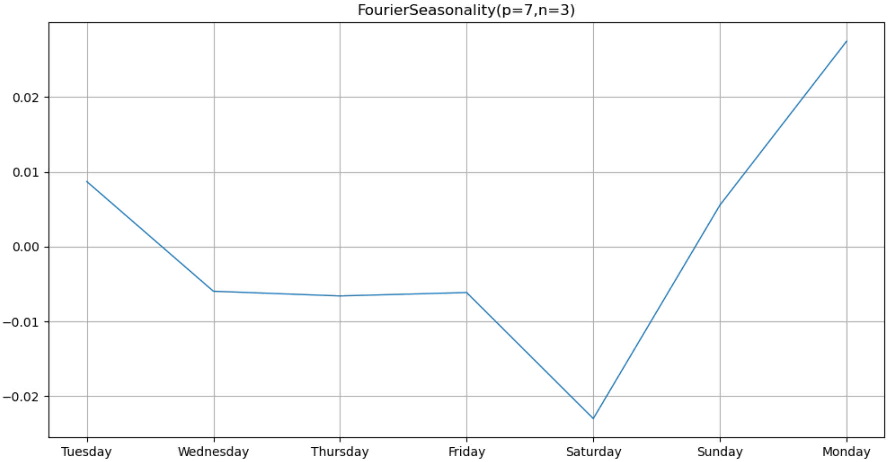
Table of Contents

5 Facebook Prophet

Visual representation of seasonalities



Visual representation of seasonalities



- Used for approximating smooth functions
- C_j - number of sine and cosine terms for period $p^{(j)}$
- $\boldsymbol{\theta}^{(j)}$ - column vector of Fourier coefficients

$$\boldsymbol{\theta}^{(j)} = \begin{pmatrix} \beta_1^{(j)} & \beta_2^{(j)} & \dots & \beta_{2C_j-1}^{(j)} & \beta_{2C_j}^{(j)} \end{pmatrix}^T$$

$$s(p^{(j)}, t; \boldsymbol{\theta}^{(j)}) = \sum_{l=1}^{C_j} (\beta_{2l-1} cos(\frac{2\pi t}{p^{(j)}}) + \beta_{2l} sin(\frac{2\pi t}{p^{(j)}}))$$

Let's vectorize this!

- $\mathbf{f}_k^{(j)} \in \mathbb{R}^{2C_j}$ - the sine and cosine terms for point $t^{(k)} \in T$

$$\mathbf{f}_k^{(j)} = \left(cos(\frac{2\pi t^{(k)}}{p^{(j)}}), sin(\frac{2\pi t^{(k)}}{p^{(j)}}), ..., cos(\frac{2\pi C_j t^{(k)}}{p^{(j)}}), sin(\frac{2\pi C_j t^{(k)}}{p^{(j)}})\right)$$

- $\mathbf{F}^{(j)} \in \mathbb{R}^{|T| \times 2C_j}$ - matrix, each row is $\mathbf{f}_k^{(j)}$
- $\mathbf{fs}^{(j)} \in \mathbb{R}^{|T|}$ - column vector representation of the Fourier seasonality at every time-point

$$\mathbf{fs}^{(j)} = \mathbf{F}^{(j)} \boldsymbol{\theta}^{(j)}$$

Table of Contents

5 Facebook Prophet

Why is the noise normal?

- The trend and seasonality model the systematic structure in the data
- The residual - unsystematic, idiosyncratic noise—measurement error, unmodeled influences, natural variability
- The sum of these independent residual factors is normal - ***CLT***
- Variants:
  - Student's t - heavy tails
  - Conditional variance - heteroscedasticity
  - Poisson, Negative Binomial - count data

A hyper-prior for noise

$$y \sim \mathcal{N}(\mathbf{g} + \sum_{j=1}^{m} \mathbf{fs}^{(j)}, \sigma)$$

- Modeling noise is a side effect of modeling the outcome
- σ is also a random variable

$$\sigma \sim \mathcal{HN}(0, 0.5)$$

The complete model

$$
\begin{aligned}
w &\sim \mathcal{N}(0, 5) \\
b &\sim \mathcal{N}(0, 5) \\
\boldsymbol{\delta} &\sim Laplace(0, 0.05) \\
\boldsymbol{\theta}^{(1)} &\sim \mathcal{N}(0, 10) \\
\boldsymbol{\theta}^{(2)} &\sim \mathcal{N}(0, 10) \\
\sigma &\sim \mathcal{HN}(0, 0.5) \\
y &\sim \mathcal{N}(\mathbf{g} + \mathbf{fs}^{(1)} + \mathbf{fs}^{(2)}, \sigma)
\end{aligned}
$$

The complete model

5 Facebook Prophet

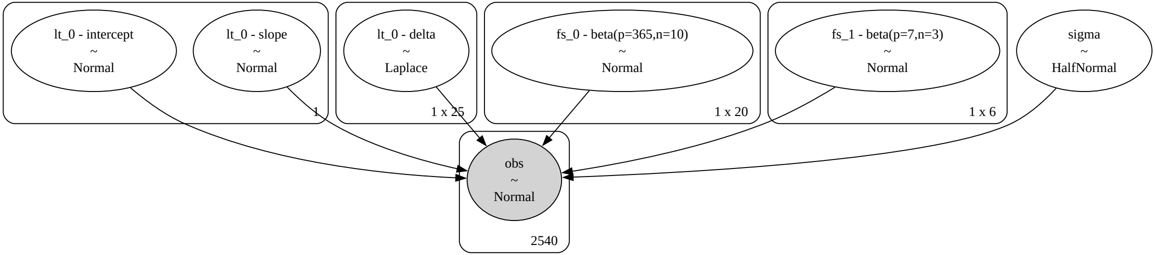


Table of Contents

6 Future research

Problem description

- Your store is selling winter equipment
- You offer a new product for three months and you need to forecast demand for the next year
- You know from your experience with the other products you sell that:
  - Your store is gaining traction (sales show an increasing trend)
  - There is a strong yearly seasonal effect - you sell more in autumn and winter
- Our problem is harder
  - Our store was only selling one product
  - Multiple new products that we have sold for three months - we need to forecast one year demand for each one of them
- How to use what you already know when modeling your 3 month samples?

The traditional Bayesian approach

- We can use Facebook Prophet with ***informative*** priors
- If we know that the trend is increasing, we can switch from $w \sim \mathcal{N}(0,5)$ to $w \sim \mathcal{N}(2,5)$
  - Or $w \sim \mathcal{N}(1,5)$?
- But what to do about the yearly seasonality?

- Let's look at the slope
- Let w be the ***shared*** slope
- Let w_i be the ***individual*** slopes

$$w \sim \mathcal{N}(0, 5)$$
$$w_i \sim \mathcal{N}(w, \frac{5}{\textbf{shrinkage}})$$

- ***Partial pooling***

Posterior $\rightarrow$ prior

- Complete Bayesian setup
  - **prior** $\rightarrow$ **posterior** $\rightarrow$ **prior** $\rightarrow \ldots$
- **prior** $\rightarrow$ **posterior**
  - Bayes theorem
- **posterior** $\rightarrow$ **prior**
  - Keep distribution shape, estimate parameters from posterior sample
  - MvN approximation of the posterior as a prior
- Fit Prophet on the large time series to get posterior
- Use posterior as prior for Prophet when fitting short time series

Why does it work?

- ***Empirical priors***
- Transfer learning + principled regularization
- The loc of the posterior - transfer learning
- The scale of the posterior - uncertainty estimate, principled regularization
- We also do standard regularization by adding ”potentials”
  - But let’s keep that discussion for another time

Combined approach

- Fit Prophet on the large time series to get posterior
- Use posterior as prior for the shared parameters in the hierarchical prophet
- Result: **shrinkage** does not need to be tuned!
- Maybe we don't even need to change the loc of the shared parameters?

Results

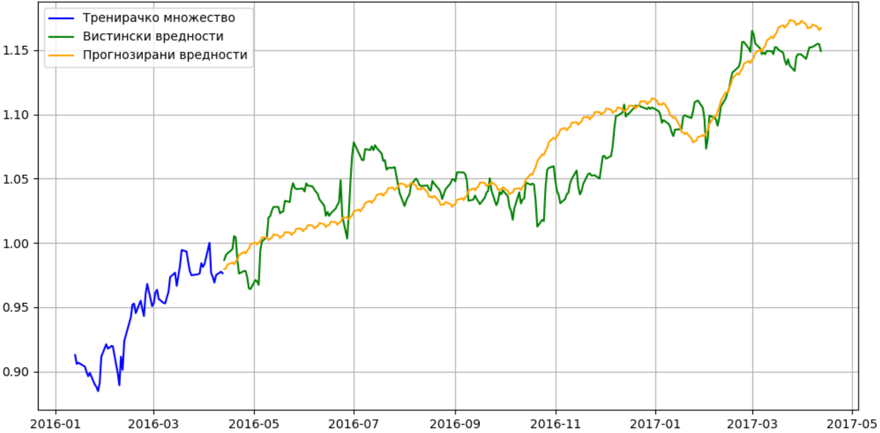
	MSE	**RMSE**	**MAE**	**MAPE**	$\downarrow$
Prophet	0.1147	0.2696	0.2335	0.2421	/
Timeseers	0.0505	0.1801	0.1535	0.1541	36.35%
Vangja hierarchical	0.0413	0.1640	0.1424	0.1384	10.19%
Vangja a posteriori	0.0380	0.1583	0.1342	0.1324	4.34%
Vangja combined	**0.0348**	**0.1493**	**0.1284**	**0.1297**	2.04%

Results

6 Future research

Results

6 Future research



Q&A

Thank you for listening!

Your feedback will be highly appreciated!