

WINTER CONFERENCE IN STATISTICS BAYESIAN MACHINE LEARNING

BAYESICS

MATTIAS VILLANI

**DEPARTMENT OF STATISTICS
STOCKHOLM UNIVERSITY**

AND

**DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE
LINKÖPING UNIVERSITY**

- **An Intro to Bayesian inference**
- **Gaussian Processes**
- **Gaussian Processes in Action**
- **Topic Models for Text**

Slides and code:

<https://github.com/mattiasvillani/WinterConfHemavan2019>

- **Bayesian inference**
- The **normal model** with known variance
- The **linear regression** model
- **Regularization priors**

- **Normal data** with **known variance**:

$$X_1, \dots, X_n | \theta \stackrel{iid}{\sim} N(\theta, \sigma^2).$$

- **Likelihood** from independent observations: x_1, \dots, x_n

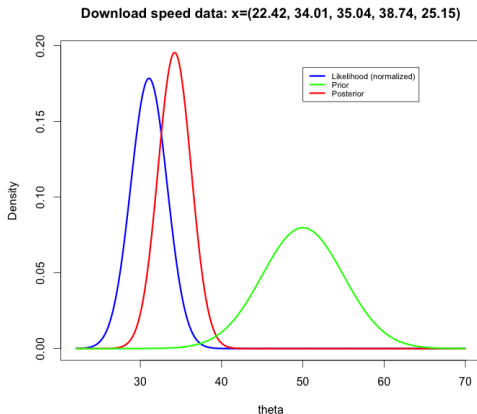
$$p(x_1, \dots, x_n | \theta) \propto \exp \left(-\frac{1}{2(\sigma^2/n)} (\theta - \bar{x})^2 \right)$$

- **Maximum likelihood**: $\hat{\theta} = \bar{x}$ maximizes $p(x_1, \dots, x_n | \theta)$.

- **Likelihood function**: $p(x_1, \dots, x_n | \theta)$ as a function of θ .

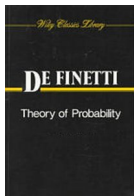
EXAMPLE: AM I REALLY GETTING MY 50MBIT/SEC?

- My broadband provider promises me at least 50Mbit/sec.
- **Data:** $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Measurement errors:** $\sigma = 5$ (± 10 Mbit with 95% probability)
- The likelihood function is proportional to $N(\bar{x}, \sigma^2/n)$ density.



UNCERTAINTY AND SUBJECTIVE PROBABILITY

- $\Pr(\theta \geq 50 | \text{data})$ only makes sense if θ is random.
- But θ may be a fixed natural constant?
- **Bayesian: doesn't matter if θ is fixed or random.**
- Do **You** know the value of θ or not?
- $p(\theta)$ reflects Your knowledge/**uncertainty** about θ .
- **Subjective probability.**
- The statement $\Pr(10\text{th decimal of } \pi = 9) = 0.1$ makes sense.



- **Bayesian learning** about a model parameter θ :
 - **prior** knowledge as a probability distribution $p(\theta)$.
 - collect data and form the **likelihood** $p(\text{Data}|\theta)$.
 - **combine prior** and **data** information.
- **How to combine** the **data** and **prior** information?
- **Bayes' theorem**

$$p(\theta|\text{Data}) = \frac{p(\text{Data}|\theta)p(\theta)}{p(\text{Data})}$$

$$p(\theta|\text{Data}) \propto p(\text{Data}|\theta)p(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \cdot \text{Prior}$$

■ Prior

$$\theta \sim N(\mu_0, \tau_0^2)$$

■ Posterior

$$\begin{aligned} p(\theta|x_1, \dots, x_n) &\propto p(x_1, \dots, x_n|\theta, \sigma^2)p(\theta) \\ &\propto N(\theta|\mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$

$$\mu_n = w\bar{X} + (1 - w)\mu_0,$$

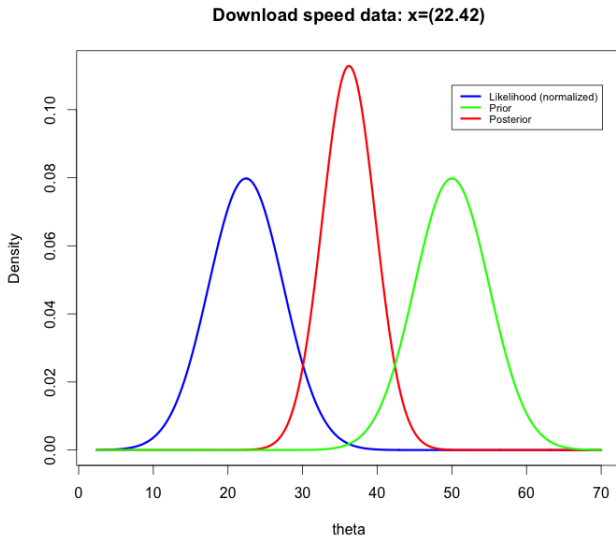
and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

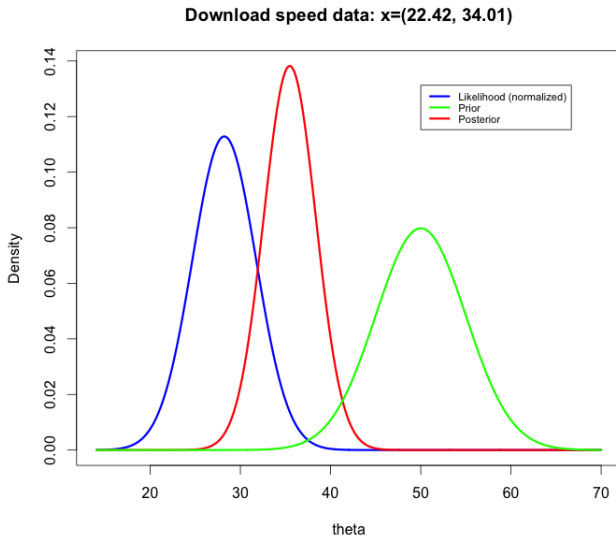
EXAMPLE: DOWNLOAD SPEED

- **Data:** $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Model:** $X_1, \dots, X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (± 10 Mbit with 95% probability)
- My **prior:** $\theta \sim N(50, 5^2)$.

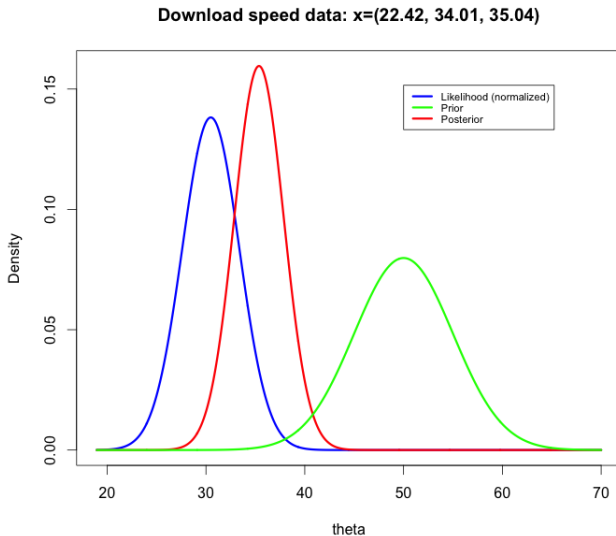
DOWNLOAD SPEED N=1



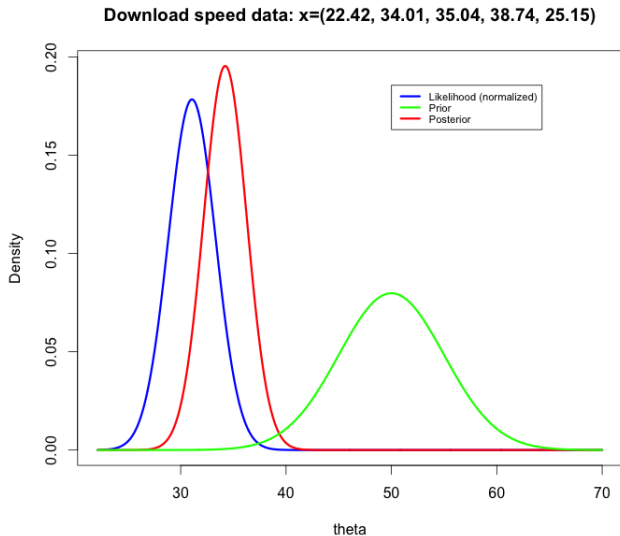
DOWNLOAD SPEED N=2



DOWNLOAD SPEED $N=3$



DOWNLOAD SPEED N=5



■ Prediction

$$p(\tilde{y}|\mathbf{y}) = \int p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta$$

■ Decision making

$$\operatorname{argmax}_{a \in \mathcal{A}} E_{p(\theta|\mathbf{y})}[U(a, \theta)]$$

■ Model inference

$\Pr(M_i|\mathbf{y})$ for models M_i in a collection of models \mathcal{M}

- **Smoothness priors** - Use extremely flexible nonlinear models and encode smoothness via the prior.

- The linear regression model in **matrix form**

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times k)(k \times 1)}{\mathbf{X}\beta} + \underset{(n \times 1)}{\varepsilon}$$

- Normal errors: $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, so $\varepsilon \sim N(0, \sigma^2 I_n)$.

- **Likelihood**

$$\mathbf{y} | \beta, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

- Standard **non-informative prior**: uniform on $(\beta, \log \sigma^2)$

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

- **Joint posterior** of β and σ^2 :

$$\begin{aligned}\beta | \sigma^2, \mathbf{y} &\sim N[\hat{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] \\ \sigma^2 | \mathbf{y} &\sim \text{Inv-}\chi^2(n-k, s^2)\end{aligned}$$

where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$.

- **Simulate** from the joint posterior by simulating from

- $p(\sigma^2 | \mathbf{y})$
- $p(\beta | \sigma^2, \mathbf{y})$

- **Marginal posterior** of β :

$$\beta | \mathbf{y} \sim t_{n-k}[\hat{\beta}, s^2 (\mathbf{X}'\mathbf{X})^{-1}]$$

■ Joint prior for β and σ^2

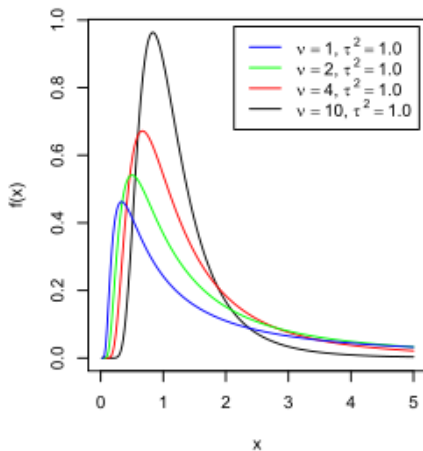
$$\begin{aligned}\beta|\sigma^2 &\sim N(\mu_0, \sigma^2 \Omega_0^{-1}) \\ \sigma^2 &\sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

■ Posterior

$$\begin{aligned}\beta|\sigma^2, \mathbf{y} &\sim N[\mu_n, \sigma^2 \Omega_n^{-1}] \\ \sigma^2|\mathbf{y} &\sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)\end{aligned}$$

$$\begin{aligned}\mu_n &= (\mathbf{X}'\mathbf{X} + \Omega_0)^{-1} (\mathbf{X}'\mathbf{X}\hat{\beta} + \Omega_0\mu_0) \\ \Omega_n &= \mathbf{X}'\mathbf{X} + \Omega_0 \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (\mathbf{y}'\mathbf{y} + \mu_0'\Omega_0\mu_0 - \mu_n'\Omega_n\mu_n)\end{aligned}$$

■ Scaled inverse χ^2 distribution



RIDGE REGRESSION = NORMAL PRIOR

- Problem: too many covariates leads to **over-fitting**.
- **Smoothness/shrinkage/regularization prior**

$$\beta_i | \sigma^2 \stackrel{iid}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Equivalent to **penalized likelihood**:

$$-2 \cdot \log p(\beta | \sigma^2, \mathbf{y}, \mathbf{X}) \propto (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta' \beta$$

- Posterior mean gives **ridge regression** estimator

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1} \mathbf{X}'\mathbf{y}$$

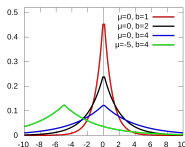
- When $\mathbf{X}'\mathbf{X} = I$. **Shrinkage**

$$\tilde{\beta} = \frac{1}{1 + \lambda} \hat{\beta}$$

LASSO REGRESSION = LAPLACE PRIOR

- **Lasso** is equivalent to posterior mode under Laplace prior

$$\beta_i | \sigma^2 \stackrel{iid}{\sim} \text{Laplace} \left(0, \frac{\sigma^2}{\lambda} \right)$$



- **Laplace prior:**
 - heavy tails
 - many β_i close to zero, but some β_i can be very large.
- **Normal prior**
 - light tails
 - all β_i 's are similar in magnitude and no β_i very large.

ESTIMATING THE SHRINKAGE

- Cross-validation is often used to determine the degree of smoothness, λ .
- Bayesian: λ is **unknown** \Rightarrow **use a prior** for λ .
- $\lambda \sim \text{Inv-}\chi^2(\eta_0, \lambda_0)$. The user specifies η_0 and λ_0 .
- Hierarchical setup:

$$\begin{aligned}\mathbf{y}|\beta, \mathbf{X} &\sim N(\mathbf{X}\beta, \sigma^2 I_n) \\ \beta|\sigma^2, \lambda &\sim N(\mathbf{0}, \sigma^2 \lambda^{-1} I_m) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \\ \lambda &\sim \text{Inv-}\chi^2(\eta_0, \lambda_0)\end{aligned}$$

- **Posterior:**

$$p(\beta, \sigma^2, \lambda|\mathbf{y}) = p(\beta|\sigma^2, \lambda, \mathbf{y})p(\sigma^2|\lambda, \mathbf{y})p(\lambda|\mathbf{y})$$

■ Polynomial regression

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k.$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where

$$\mathbf{X} = (1, x, x^2, \dots, x^k).$$

■ Problem: higher order polynomials can overfit the data.

■ Solution: shrink higher order coefficients harder:

$$\beta | \sigma^2 \sim N \left[\mathbf{0}, \begin{pmatrix} 100 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{2\lambda} & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & \frac{1}{k\lambda} \end{pmatrix} \right]$$

- Quadratic relationship between pain relief (y) and time (x)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

- At what time x_{max} is there **maximal pain relief**?

$$x_{max} = -\beta_1 / 2\beta_2$$

.

- Posterior distribution of x_{max} can be obtained by change of variable. Cauchy-like.
- Easy to obtain marginal posterior $p(x_{max} | \mathbf{y}, \mathbf{X})$ by **simulation**:
 - Simulate N coefficient vectors from the posterior $\beta, \sigma^2 | \mathbf{y}, \mathbf{X}$
 - For each simulated β , compute $x_{max} = -\beta_1 / 2\beta_2$.
 - Plot a histogram. Converges to $p(x_{max} | \mathbf{y}, \mathbf{X})$ as $N \rightarrow \infty$.

FINDING THE TIME FOR MAXIMUM

