# Winter Conference in Statistics
# Bayesian Machine Learning

## Bayesics


Mattias Villani

**Department of Statistics**
**Stockholm University**
**and**
**Department of Computer and Information Science**
**Linköping University**

# Lecture overview

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

**Slides and code**:
https://github.com/mattiasvillani/WinterConfHemavan2019

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

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

- **Likelihood** from independent observations: $x_1, ..., x_n$

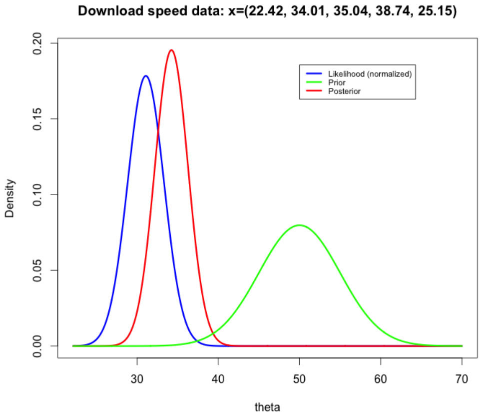
$$p(x_1, ..., x_n|\theta) \propto \exp\left(-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right)$$

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

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

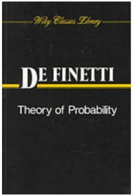
# Example: Am I really getting my 50Mbit/sec?

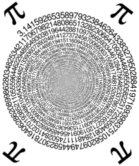
- My broadband provider promises me at least 50Mbit/sec.
- **Data**: $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Measurement errors**: $\sigma = 5$ ($\pm$10Mbit with 95% probability)
- The likelihood function is proportional to $\mathrm{N}(\bar{x}, \sigma^2/n)$ density.



# Uncertainty and subjective probability

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



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

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

$$p(\theta|Data) \propto p(Data|\theta)p(\theta)$$

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

- **Prior**

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

- **Posterior**

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

where

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

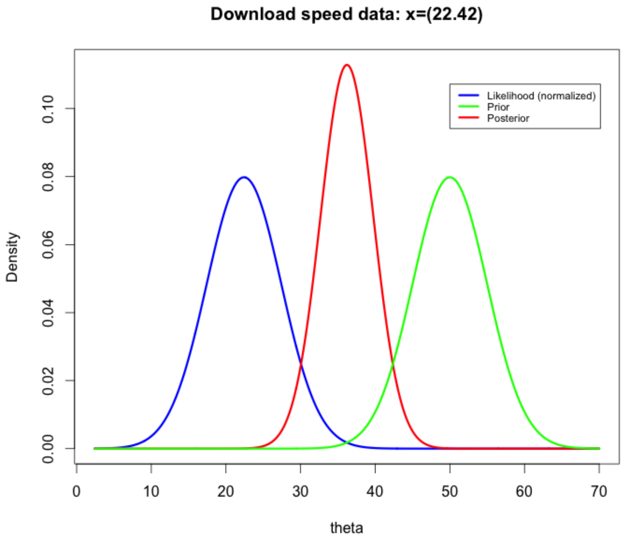
$$\mu_n = w\bar{x} + (1 - w)\mu_0,$$

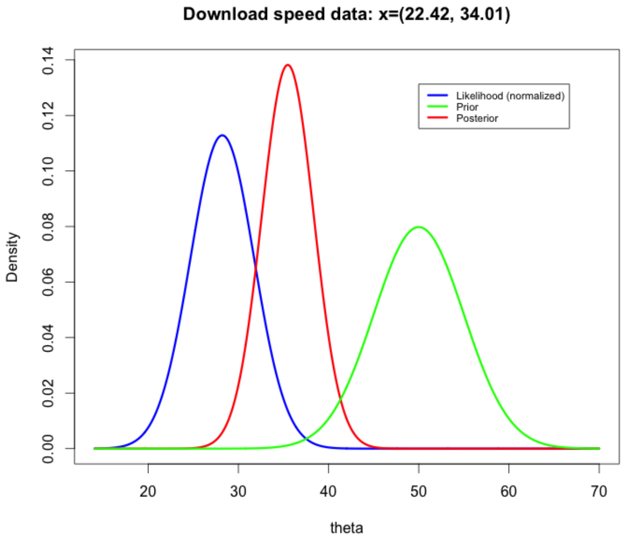
and

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

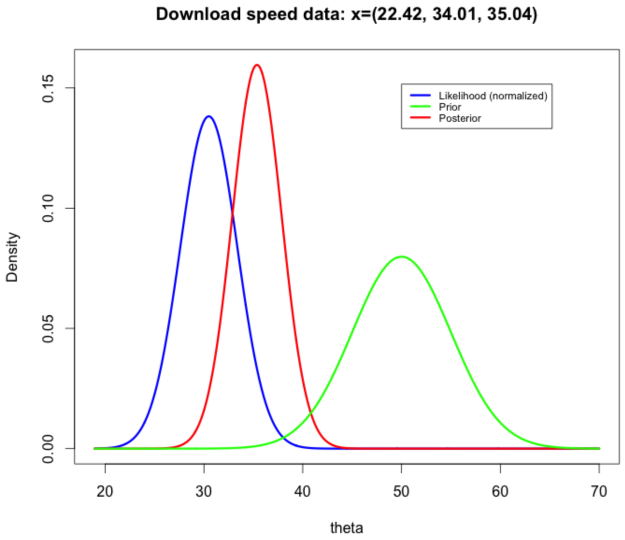
- **Data**: $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Model**: $X_1, \ldots, X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ ($\pm$10Mbit with 95% probability)
- My **prior**: $\theta \sim N(50, 5^2)$.

# Download speed n=1

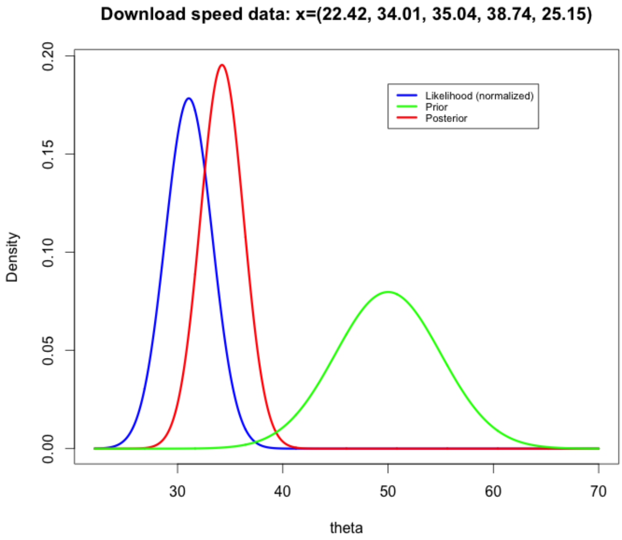



Download speed data: x=(22.42, 34.01)
Likelihood (normalized)
Prior
Posterior
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
20
30
40
50
60
70
theta

# Download speed n=3



# Download speed n=5



# Why Bayesian inference in Machine Learning

- **Prediction**

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

- **Decision making**

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

- **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: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, 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 $\beta$ and $\sigma^2$:

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

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 $\beta$ :

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

- **Joint prior** for $\beta$ and $\sigma^2$

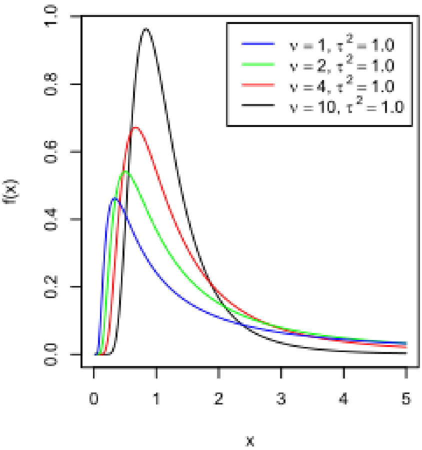
$$\begin{aligned}\beta|\sigma^2 &\sim N\left(\mu_0, \sigma^2\Omega_0^{-1}\right)\\ \sigma^2 &\sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)\end{aligned}$$

- **Posterior**

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

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

- **Scaled inverse $\chi^2$ distribution**



# Ridge regression = normal prior

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

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

- Equivalent to **penalized likelihood**:

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

- Posterior mean gives **ridge regression** estimator

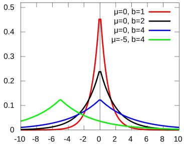
$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}'\mathbf{y}$$

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

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

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

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




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

# Estimating the shrinkage

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

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

- **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**

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

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

where

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

- Problem: higher order polynomials can overfit the data.
- Solution: shrink higher order coefficients harder:

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

- 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 $\beta$, compute $x_{max} = -\beta_1/2\beta_2$.
  - Plot a histogram. Converges to $p(x_{max}|\mathbf{y}, \mathbf{X})$ as $N \to \infty$.

# Finding the time for maximum