Bayesian Linear Regression

Guest lecture at KTH

Mattias Villani

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

- **Bayesian inference**
- The **Normal model** with known variance
- **Normal model** with conjugate prior
- The **linear regression** model
- **Regularization priors**

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

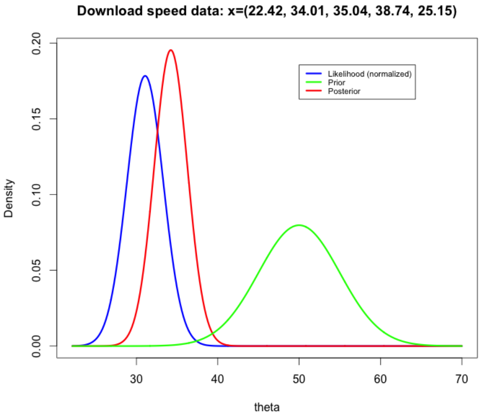
$$X_1, ..., X_n|\theta \overset{iid}{\sim} \mathrm{N}(\theta, \sigma^2).$$

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

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

- **Maximum likelihood**: $\hat{\theta} = \bar{x}$ maximizes $p(x_1, ..., x_n|\theta)$.
- Given the data $x_1, ..., x_n$, plot $p(x_1, ..., x_n|\theta)$ as a function of θ.

- 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.

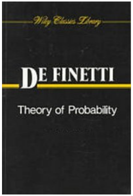


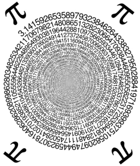
- Say it out loud:

  The likelihood function is
  the probability of the observed data
  considered as a function of the parameter.

- Likelihood function is **NOT** a probability distribution for θ.
- Statements like $\Pr(\theta > c)$ makes no sense.
- Unless ...

- $\Pr(\theta < 0.6|\text{data})$ 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 $\Pr(\text{10th decimal of } \pi = 9) = 0.1$ makes sense.





- **Bayesian learning** about a model parameter θ:
  - state your **prior** knowledge as a probability distribution $p(\theta)$.
  - collect **data** $\mathbf{x}$ and form the **likelihood** function $p(\mathbf{x}|\theta)$.
  - **combine** prior knowledge $p(\theta)$ with data information $p(\mathbf{x}|\theta)$.
- **How to combine** the two sources of information?

Bayes' theorem

- How to **update** from **prior** $p(\theta)$ to **posterior** $p(\theta|Data)$?
- **Bayes' theorem** for events A and B

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

- Bayes' Theorem for a model parameter θ

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

- It is the prior $p(\theta)$ that takes us from $p(Data|\theta)$ to $p(\theta|Data)$.
- A probability distribution for θ is extremely useful. **Predictions**. **Decision making**.

- Bayes theorem

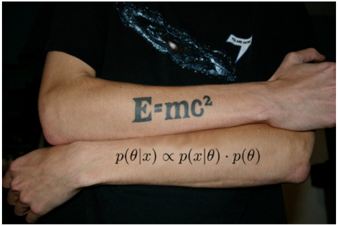
$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}$$

- All you need to know:

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

or

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



- **Model**

$$x_1, \ldots, x_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$

- **Prior**

$$p(\theta) \propto c \text{ (a constant)}$$

- **Likelihood**

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

- **Posterior**

$$\theta | x_1, \ldots, x_n \sim N(\bar{x}, \sigma^2/n)$$

- **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

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

- Proof: complete the squares in the exponential.

$$\theta \sim N(\mu_0, \tau_0^2) \overset{x_1,\ldots,x_n}{\Longrightarrow} \theta | x \sim N(\mu_n, \tau_n^2).$$

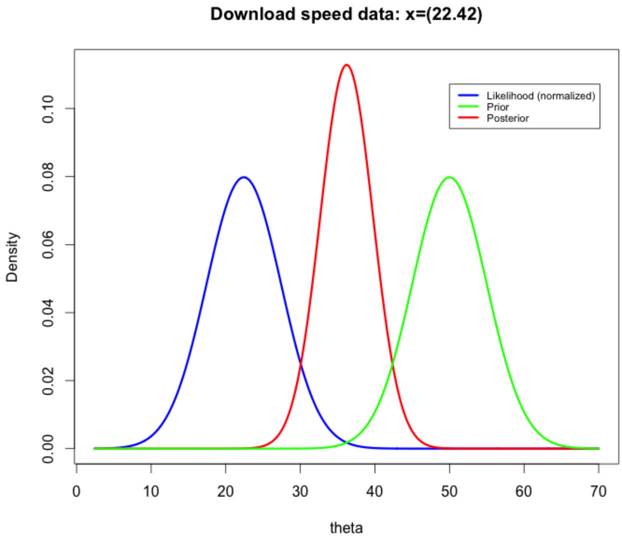
Posterior precision = Data precision + Prior precision

Posterior mean =

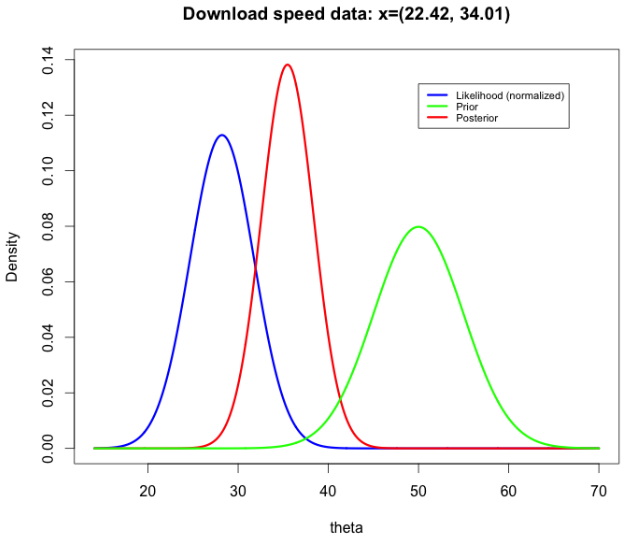
$$\frac{\text{Data precision}}{\text{Posterior precision}}(\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}}(\text{Prior mean})$$

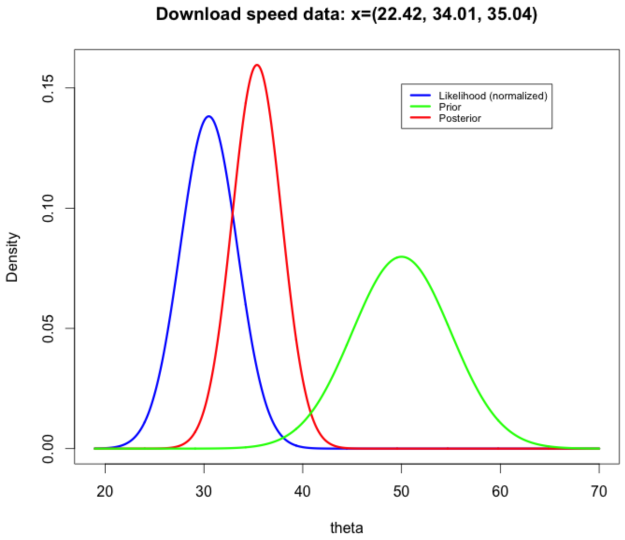
- **Data**: $x = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- **Model**: $X_1, ..., X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (measurements can vary $\pm$10MBit with 95% probability)
- My **prior**: $\theta \sim N(50, 5^2)$.

Download speed n=1

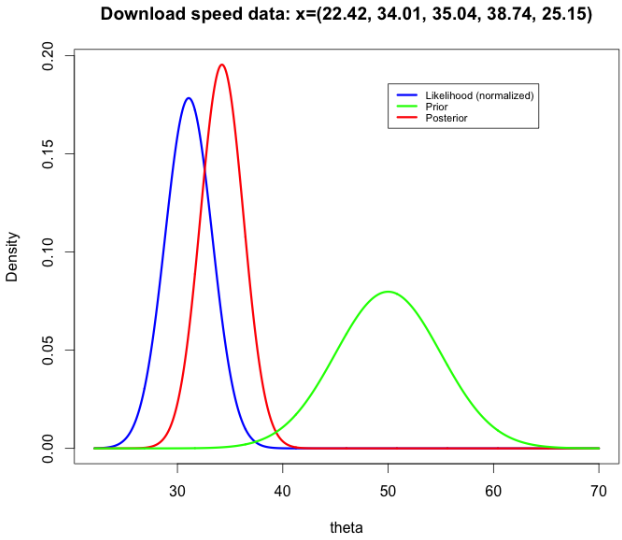


Download speed n=2




Download speed data: x=(22.42, 34.01, 35.04)
Likelihood (normalized)
Prior
Posterior
Density
0.00
0.05
0.10
0.15
theta
20
30
40
50
60
70

Download speed n=5



- **Model**

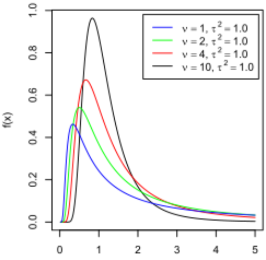
$$y_1, \ldots, y_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **Conjugate prior**

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

- **Scaled inverse χ^2 distribution**



- **Posterior**

$$\theta|\mathbf{y},\sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|\mathbf{y} \sim Inv\text{-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

- **Marginal posterior**

$$\theta|\mathbf{y} \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

- The ordinary **linear regression** model:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ & \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

- Parameters $\theta = (\beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$.

- **Assumptions**:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ (linear function)
  - $Var(y_i) = \sigma^2$ (homoscedasticity)
  - $Corr(y_i, y_j | X, \beta, \sigma^2) = 0$, $i \neq j$.
  - Normality of ε_i.
  - The x's are assumed known (non-random).

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

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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Usually $x_{i1} = 1$, for all i. β_1 is the intercept.
- **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:

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

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 β :

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

- **Joint prior** for β and σ^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**

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

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

- **Polynomial regression**

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

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

where

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

Ridge regression = normal prior

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

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

- Larger λ gives smoother fit. Note: $\Omega_0 = \lambda I$.
- 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

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

- **Shrinkage** toward zero

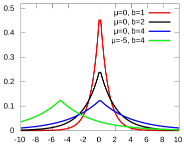
$$\text{As } \lambda \to \infty,\ \tilde{\beta} \to 0$$

- When $\mathbf{X}'\mathbf{X} = I$

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

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

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



- The **Bayesian shrinkage** prior is **interpretable**. **Not ad hoc**.
- Laplace distribution have heavy tails.
- **Laplace prior**: many β_i close to zero, but some β_i very large.
- Normal distribution have light tails.
- **Normal prior**: all β_i's are similar in magnitude.

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

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

so $\Omega_0 = \lambda I_m$.

- The **joint posterior** of β, σ^2 and λ is

$$
\begin{aligned}
\beta|\sigma^2, \lambda, \mathbf{y} &\sim N\left(\mu_n, \Omega_n^{-1}\right) \\
\sigma^2|\lambda, \mathbf{y} &\sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right) \\
p(\lambda|\mathbf{y}) &\propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}^T\mathbf{X}+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)
\end{aligned}
$$

where $\Omega_0 = \lambda I_m$, and $p(\lambda)$ is the prior for λ, and

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

- **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 β, 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