Workshop: Intro to Bayesian Learning
Lecture 4 - Bayesian Classification and Posterior Approximation

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com   @matvil   @matvil   mattiasvillani

Overview

- **Bayesian logistic regression**
- **Posterior approximation**

Binary regression

- **Logistic regression**

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- **Probit regression**

$$\Pr(y_i = 1 | \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression

$$\Pr(y_i = c \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_c)}{\sum_{k=1}^{C} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}$$

- **Likelihood**

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\left(\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right)^{y_i}}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- Problem: no conjugate prior. **Posterior is intractable**. Now what?

Likelihood information

- **Observed information** in likelihood $\ln p(\mathbf{x}|\theta)$ for **given** data $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$

$$J_{\mathbf{x}}(\hat{\theta}) = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}$$

where $\hat{\theta}$ is the maximum likelihood estimate.

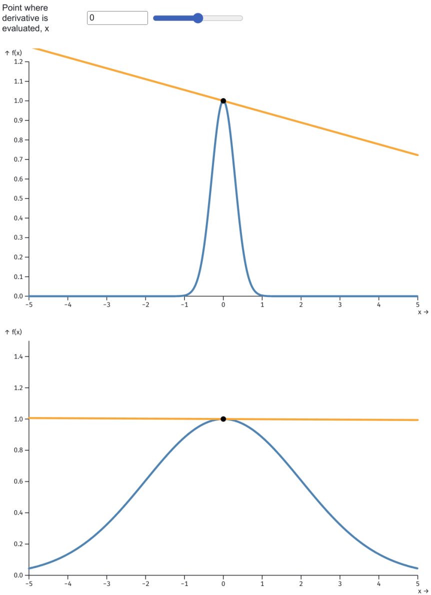
- Multiparameter **observed information matrix**

$$J_{\mathbf{x}}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

- Example: $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$

$$\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2^2} \end{pmatrix}$$

ʘʘ Second derivative measures curvature



Posterior asymptotics

Normal posterior approximation

The posterior can in large samples be approximated by

$$\boldsymbol{\theta}|\mathbf{x} \overset{\mathrm{a}}{\sim} \mathrm{N}\left(\tilde{\boldsymbol{\theta}}, J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})\right)$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode and

$$J_{\mathbf{x}}(\tilde{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

is the $d \times d$ observed posterior information matrix at $\tilde{\boldsymbol{\theta}}$.

- Important: sufficient with proportional form

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Example: gamma posterior

- **Poisson model**: $\theta|y_1, ..., y_n \sim \mathrm{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$

$$\log p(\theta|y_1, ..., y_n) \propto (\alpha + \sum_{i=1}^{n} y_i - 1) \log \theta - \theta(\beta + n)$$

- First derivative of log density

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\theta} - (\beta + n)$$

$$\tilde{\theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}$$

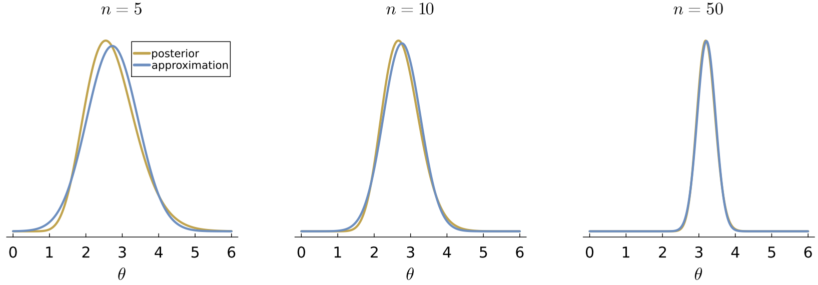
- Second derivative at mode $\tilde{\theta}$

$$\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial \theta^2}\Big|_{\theta=\tilde{\theta}} = -\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\left(\frac{\alpha+\sum_{i=1}^{n} y_i - 1}{\beta+n}\right)^2} = -\frac{(\beta + n)^2}{\alpha + \sum_{i=1}^{n} y_i - 1}$$

- **Normal approximation**

$$N\left[\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}, \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{(\beta + n)^2}\right]$$

Example: gamma posterior for eBay bidders data

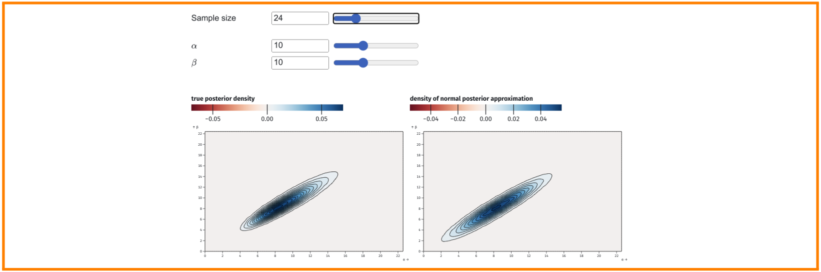


Numerical Normal approximation of posterior

- Standard **optimization routines** may be used to obtain the normal approximation (e.g. `optim in` R).
  - **Input**: expression proportional to $\log p(\boldsymbol{\theta}|\mathbf{x})$. Initial values.
  - **Output**: $\log p(\tilde{\boldsymbol{\theta}}|\mathbf{x})$, $\tilde{\boldsymbol{\theta}}$ and Hessian matrix $(-J_{\mathbf{x}}(\tilde{\boldsymbol{\theta}}))$.
- **Automatic differentation** - efficient derivatives on computer.
- **Re-parametrization** may improve normal approximation:
  - If $\theta \geq 0$ use $\phi = \log(\theta)$.
  - If $0 \leq \theta \leq 1$, use $\phi = \ln[\theta/(1-\theta)]$.
  - Don't forget the **Jacobian**!
- Posterior approximation of functions $g(\boldsymbol{\theta})$ by simulation from

$$\boldsymbol{\theta}|\mathbf{y} \overset{a}{\sim} N\left(\tilde{\boldsymbol{\theta}}, J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})\right)$$

Normal approx of posterior in Beta regression



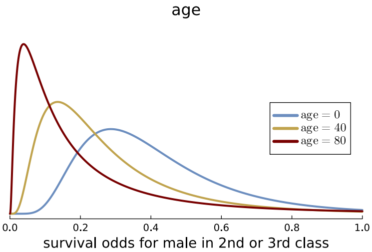
Logistic regression - who survived the Titanic?

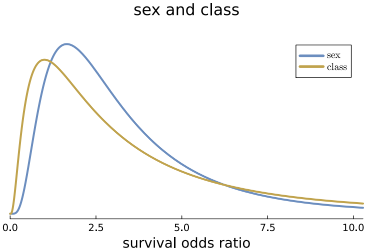
- Prior

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$$

with

$$\boldsymbol{\mu} = \big(-1, -1/80, 1, 1\big)^\top \qquad \boldsymbol{\Omega} = \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1/(80^2) & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



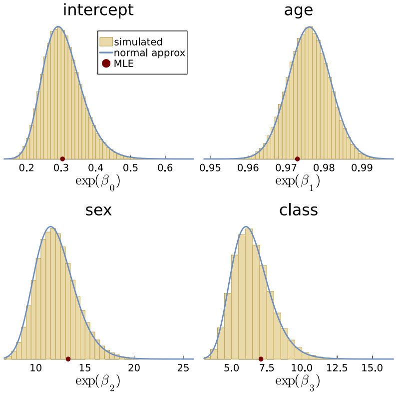


Logistic regression - who survived the Titanic?

- **Normal posterior approximation**

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\beta}})\right).$$

- Means that the posterior of each β_j is univariate normal.
- Marginal posterior for each $\exp(\beta_j)$ **is LogNormal**.



Logistic regression - who survived the Titanic?

- Comparison with non-informative prior $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^2 \boldsymbol{I}_p)$.