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**

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

- **Probit regression**

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

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

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

- **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.

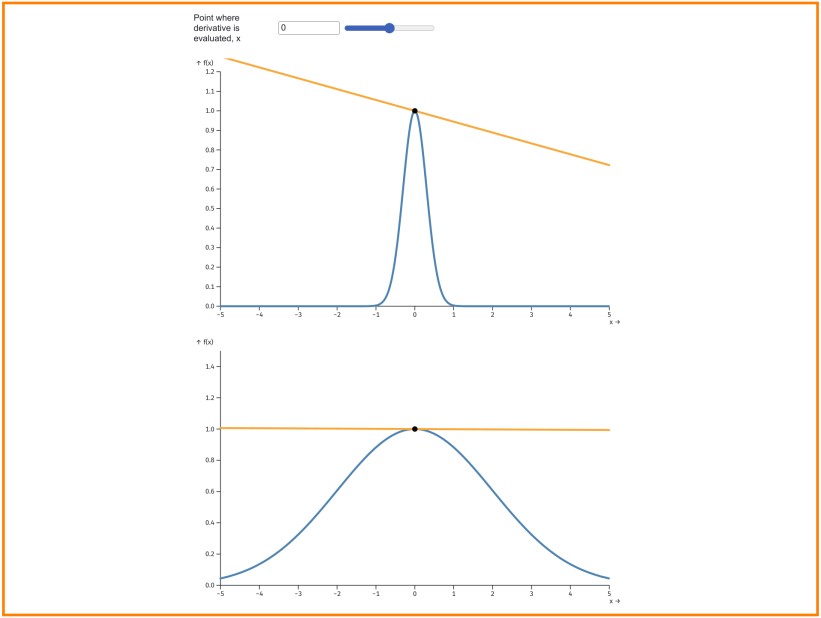
- Multiparameter **observed information matrix**

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

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

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

Second derivative measures curvature



Posterior asymptotics

Normal posterior approximation

The posterior can in large samples be approximated by

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

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

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

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

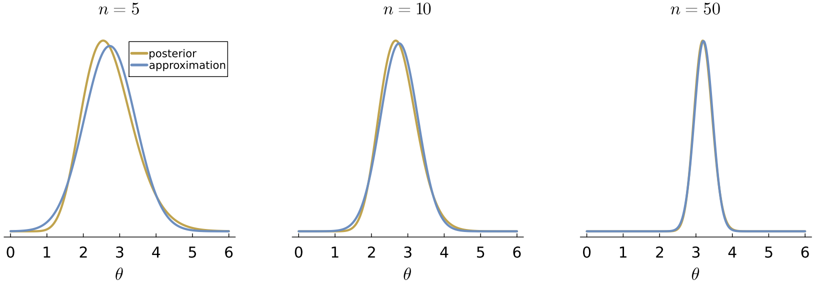
- Second derivative at mode $\tilde{\theta}$

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

- **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 **numerical optimization** (e.g. `optim in` R).
  - **Input**: function computing $\log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$. Initial values.
  - **Output**: $\tilde{\boldsymbol{\theta}}$ 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 $\phi = \log[\theta/(1-\theta)]$.
  - Don't forget the **Jacobian**!
- Posterior approximation of functions $g(\boldsymbol{\theta})$ by simulation from

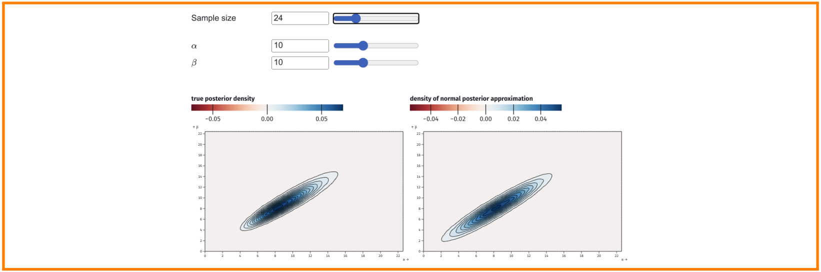
$$\boldsymbol{\theta}|\mathbf{y} \overset{\text{a}}{\sim} N\left(\tilde{\boldsymbol{\theta}}, J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})\right)$$

Normal posterior approx for logistic regression - Julia

```julia
# 0. Loading packages
using Plots, Distributions, GLM, LinearAlgebra, Optim, ForwardDiff

# 1. Setting up the log posterior function
function logisticreg_logpost(β, y, X, μ, Σ)
    loglik = sum( y.*(X*β)  .- log.(1 .+ exp.(X*β)) )
    logprior = logpdf(MvNormal(μ, Σ), β)
    return(loglik + logprior)
end

# 2. Load data
y, X = load(data)
n, p = size(X)

# 3. Set up prior
μ = zeros(p)
Σ = 10*I(p)

# 4. Initial value for the optimization
glmfit = glm(X, y, Bernoulli(), LogitLink()) # find MLE.
βₒ = coef(glmfit) # initial values from MLE.

# 5. Run optimizer with autodiff to find mode and Hessian.
optres = maximize(β -> logisticreg_logpost(β, y, X, μ, Σ), βₒ, autodiff = :forward)
βmode = Optim.maximizer(optres)

# 6. Compute Hessian to get posterior covariance matrix approximation
H(β) = ForwardDiff.hessian(β -> logisticreg_logpost(β, y, X, μ, Σ), β)
Ωᵦ = Symmetric(-inv(H(βmode))) # This is J^{-1}

# 7. Simulate from normal posterior approximation and compute odds ratios
βsim = rand(MvNormal(βmode,Ωᵦ), 10000)'
oddsratio = exp.(βsim) # 10000 × 4 matrix with draws of exp(βⱼ) in jth column.
```

- Bayesian logistic regression in **R**: R notebook.

Normal approx of posterior in Beta regression



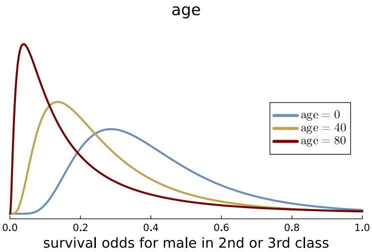
Logistic regression - who survived the Titanic?

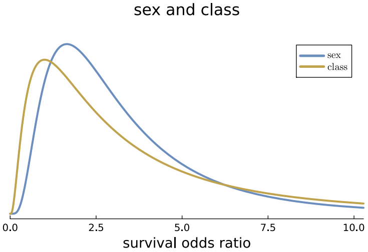
- Prior

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

with

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



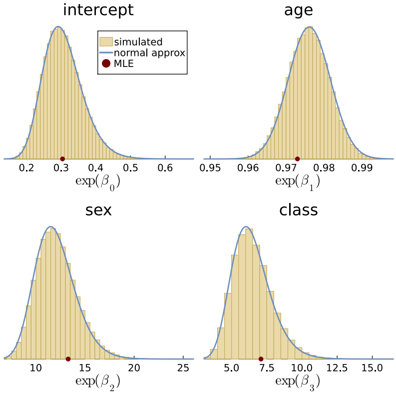


Logistic regression - who survived the Titanic?

- **Normal posterior approximation**

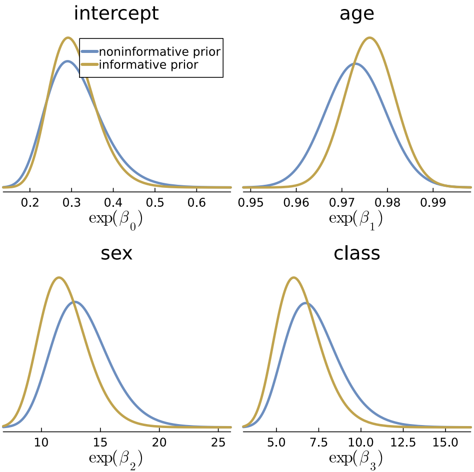
$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\beta}})\right).$$

- 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 $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^2 \boldsymbol{I_p})$.



Bayesian model comparison

- **Posterior model probabilities**

$$\underbrace{\Pr(M_k|\mathbf{y})}_{\text{posterior model prob.}} \propto \underbrace{p(\mathbf{y}|M_k)}_{\text{marginal likelihood}} \cdot \underbrace{\Pr(M_k)}_{\text{prior model prob.}}$$

- The **marginal likelihood** for model M_k with parameters θ_k

$$\underbrace{p(\mathbf{y}|M_k)} = \int p(\boldsymbol{y}|\theta_k, M_k)p(\theta_k|M_k)d\theta_k.$$

- θ_k is 'removed' by the averaging wrt prior. **Priors matter!**
- The **Bayes factor**

$$B_{12}(\boldsymbol{y}) = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Laplace approximation of marginal likelihood

- **The Laplace approximation**:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \ln p(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \ln \left| J_{\mathbf{y}}^{-1}(\hat{\boldsymbol{\theta}}) \right| + \frac{p}{2} \ln(2\pi),$$

where p is the number of unrestricted parameters.

- The **BIC approximation** assumes that $J_{\mathbf{y}}(\hat{\boldsymbol{\theta}})$ behaves like $n \cdot I_p$ in large samples and the small term $\frac{p}{2}\ln(2\pi)$ is ignored

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{p}{2} \ln n.$$

- **Log predictive score** is like a marginal likelihood, but the prior is replaced by posterior from a (small) training data. **More robust** to the prior.