Workshop: Intro to Bayesian Learning
Lecture 3 - Bayesian Regression and Regularization

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Bayesian linear regression**
- **Regularization priors**

Linear regression

- Linear Gaussian regression

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

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

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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

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

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

Posterior in linear regression - uniform prior

Gaussian linear regression with non-informative prior

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$

Prior: $p(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$

Posterior:

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$
$$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv}\!-\!\chi^2(n-p, s^2)$$

$\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
$s^2 \equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$

Marginal posterior:

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

Linear regression - conjugate prior

- **Joint prior** for $\boldsymbol{\beta}$ and σ^2

$$
\begin{aligned}
\boldsymbol{\beta}|\sigma^2 &\sim N\left(\mu_0, \sigma^2 \boldsymbol{\Omega}_0^{-1}\right) \\
\sigma^2 &\sim \mathrm{Inv-}\chi^2\left(\nu_0, \sigma_0^2\right)
\end{aligned}
$$

- Common choices:
  - $\boldsymbol{\Omega}_0 = \kappa \boldsymbol{I}_p$ (**Ridge**)
  - $\boldsymbol{\Omega}_0 = \frac{\kappa}{n} \boldsymbol{X}^\top \boldsymbol{X}$ (**Zellner's prior**).
  - $\boldsymbol{\Omega}_0 = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}$ (Noninformative **Unit information prior**)

Posterior in linear regression - conjugate prior

Gaussian linear regression with conjugate prior

Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 I_n)$

Prior:

$$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Omega}_0^{-1})$$

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

Posterior:

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_n, \sigma^2 \boldsymbol{\Omega}_n^{-1})$$

$$\sigma^2|\mathbf{y}, \mathbf{X} \sim \mathrm{Inv-}\chi^2(\nu_n, \sigma_n^2)$$

$\boldsymbol{\Omega}_n = \mathbf{X}^\top \mathbf{X} + \boldsymbol{\Omega}_0,$
$\boldsymbol{\mu}_n = \left(\boldsymbol{\Omega}_n^{-1} \mathbf{X}^\top \mathbf{X}\right)\hat{\boldsymbol{\beta}} + \boldsymbol{\Omega}_n^{-1} \boldsymbol{\Omega}_0 \boldsymbol{\mu}_0$
$\nu_n = \nu_0 + n$

Marginal posterior: $\boldsymbol{\beta}|\mathbf{y} \sim t_{\nu_n}\left(\boldsymbol{\mu}_n, \sigma_n^2 \boldsymbol{\Omega}_n^{-1}\right)$

Julia code for linear regression - conjugate prior

```julia
function BayesLinReg(y::Vector, X, μₒ, Ωₒ, vₒ, σ²ₒ, nSim)

    # Define ScaledInverseChiSquare distribution
    ScaledInverseChiSq(v,τ²) = InverseGamma(v/2,v*τ²/2)

    # Compute posterior hyperparameters
    n = length(y)
    p = size(X,2)
    XX = X'*X
    β̂ = X \ y
    Ωₙ = Symmetric(XX + Ωₒ)
    μₙ = Ωₙ\(XX*β̂ + Ωₒ*μₒ)
    vₙ = vₒ + n
    σ²ₙ = ( vₒ*σ²ₒ + (y-X*β̂)'*(y-X*β̂) + (μₙ-β̂)'*XX*(μₙ-β̂) +
        (μₙ-μₒ)'*Ωₒ*(μₙ-μₒ) )/vₙ
    invΩₙ = inv(Ωₙ)

    # Sampling from posterior
    σ²sim = zeros(nSim)
    βsim = zeros(nSim,p)
    for i ∈ 1:nSim
        # Simulate from p(σ²|y,X)
        σ² = rand(ScaledInverseChiSq(vₙ,σ²ₙ))
        σ²sim[i] = σ²

        # Simulate from p(β|σ²,y,X)
        β = rand(MvNormal(μₙ,σ²*invΩₙ))
        βsim[i,:] = β'
    end

    return μₙ, Ωₙ, vₙ, σ²ₙ, βsim, σ²sim

end
```

R for linear regression - conjugate prior

```r
# Function to simulate from the scaled inverse Chi-square distribution
rScaledInvChi2 <- function(n, df, scale){
  return((df*scale)/rchisq(n,df=df))
}

BayesLinReg <- function(y, X, mu_0, Omega_0, v_0, sigma2_0, nIter){

  # Compute posterior hyperparameters
  n = length(y) # Number of observations
  nCovs = dim(X)[2] # Number of covariates
  XX = t(X)%*%X
  betaHat <- solve(XX,t(X)%*%y)
  Omega_n = XX + Omega_0
  mu_n = solve(Omega_n,XX%*%betaHat+Omega_0%*%mu_0)
  v_n = v_0 + n
  sigma2_n = as.numeric((v_0*sigma2_0 + ( t(y)%*%y + t(mu_0)%*%Omega_0%*%mu_0 -
                                          t(mu_n)%*%Omega_n%*%mu_n))/v_n)
  invOmega_n = solve(Omega_n)

  # The actual sampling
  sigma2Sample = rep(NA, nIter)
  betaSample = matrix(NA, nIter, nCovs)
  for (i in 1:nIter){

    # Simulate from p(sigma2 | y, X)
    sigma2 = rScaledInvChi2(n=1, df = v_n, scale = sigma2_n)
    sigma2Sample[i] = sigma2

    # Simulate from p(beta | sigma2, y, X)
    beta_ = rmvnorm(n=1, mean = mu_n, sigma = sigma2*invOmega_n)
    betaSample[i,] = beta_

  }
  return(results = list(sigma2Sample = sigma2Sample, betaSample=betaSample))
}
```

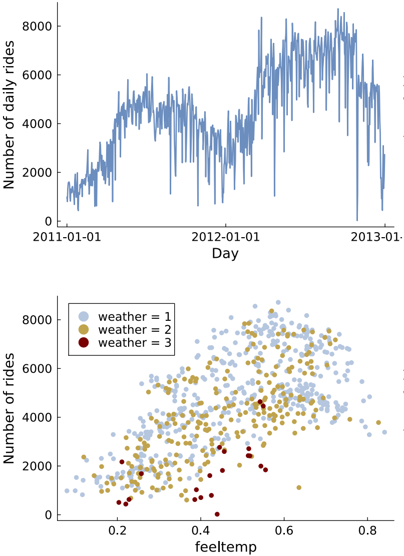
Bike share data

- **Bike share data**. Predict the number of bike rides.
- Response variable: number of rides on 731 days.

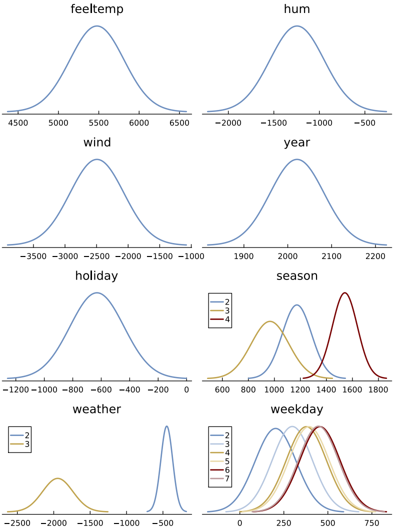
variable	description	type	values	comment
`nrides`	# of rides	counts	$\{0, 1, ...\}$	min$= 22$, max$= 8714$
`feeltemp`	perceived temp	cont.	$[0, 1]$	min$= 0.07$, max$= 0.85$
`hum`	humidity	cont.	$[0, 1]$	min$= 0.00$, max$= 0.98$
`wind`	wind speed	cont.	$[0, 1]$	min$= 0.02$, max$= 0.51$
`year`	year	binary	$\{0, 1\}$	year 2011 $= 0$
`season`	season	cat.	$\{1, 2, 3, 4\}$	winter $\rightarrow$ fall
`weather`	weather	ordinal	$\{1, 2, 3\}$	clear $\rightarrow$ rain/snow
`weekday`	day of week	cat.	$\{0, ..., 6\}$	sunday $\rightarrow$ saturday
`holiday`	holiday	binary	$\{0, 1\}$	holiday $= 1$

- Prior:
  - $\boldsymbol{\mu}_0 = (1000, 0, \ldots, 0)^\top$
  - $\boldsymbol{\Omega}_0 = \frac{\kappa_0}{n} \boldsymbol{X}^\top \boldsymbol{X}$ with $\kappa_0 = 1$ (unit information prior)
  - $\sigma_0^2 = 1000^2$ and $\nu_0 = 5$.

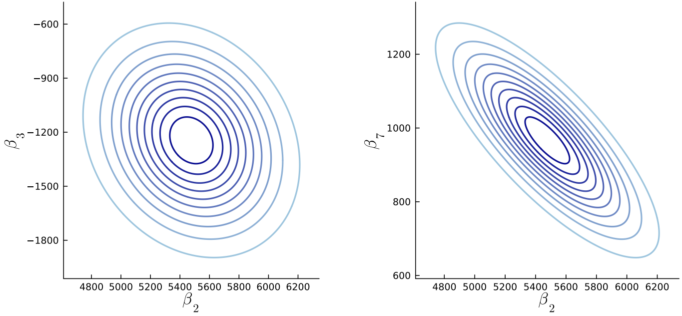
Bike share data



Bike share data - marginal posteriors of β



Bike share data - joint posteriors of β



Interactive - Bayesian regression

Ridge regression = iid normal prior

- **Shrinkage/regularization prior** $[\boldsymbol{\Omega}_0 = \lambda \boldsymbol{I}_p]$

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

- Posterior mean is the **ridge regression** estimator

$$\boldsymbol{\mu}_n = \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

- **Shrinkage** toward zero

$$\text{As } \lambda \to \infty,\ \boldsymbol{\mu}_n \to \mathbf{0}$$

- When $\mathbf{X}^\top \mathbf{X} = \boldsymbol{I}_p$

$$\boldsymbol{\mu}_n = (1 - \phi)\hat{\boldsymbol{\beta}}, \qquad \text{for } \phi = \frac{\lambda}{1 + \lambda}$$

- **Shrinkage factor** $\phi \in [0, 1]$.

Lasso regression = Laplace prior

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

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

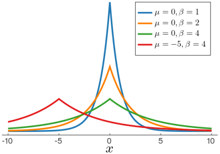
Laplace distribution

$X \sim \text{Laplace}(\mu, \beta)$ for $X \in \mathbb{R}$.

$$p(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right)$$

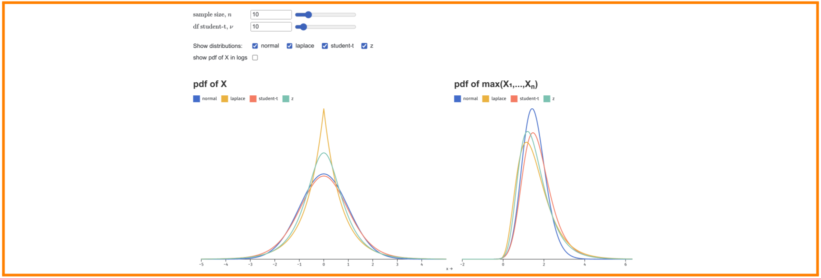
$\mathbb{E}(X) = \mu$

$\mathbb{V}(X) = 2\beta^2$



- **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.

Interactive - tails of distributions



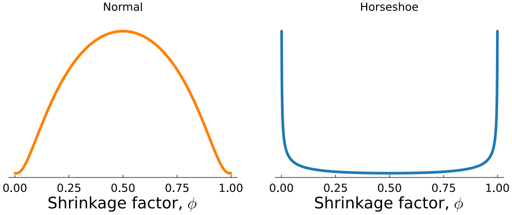
Horseshoe prior

- Normal and Laplace - one **global shrinkage** parameter λ.
- **Global-Local shrinkage**: global + local shrinkage for each β_j.
- **Horseshoe prior**:

$$\beta_j|\lambda_j^2, \tau^2 \overset{\text{ind}}{\sim} N\left(0, \sigma^2\tau^2\lambda_j^2\right)$$
$$\lambda_j \sim C^+(0,1)$$
$$\tau \sim C^+(0,1)$$

- The posterior mean for $\boldsymbol{\beta}$ satisfies approximately

$$\mu_{n,j} \approx (1-\phi_j)\hat{\beta}_j$$



Variable selection by spike-and-slab prior

- **Spike-and-slab prior**

$$\beta_j|\sigma^2, \tau^2, z_j \sim \begin{cases} 0 & \text{if } z_j = 0 \\ N\left(0, \sigma^2\tau^2\right) & \text{if } z_j = 1 \end{cases}$$

- Prior for the **variable selection indicators**

$$z_j \overset{iid}{\sim} \mathrm{Bernoulli}(\omega)$$

- This is a **mixture prior** for the β_j

$$p(\beta_j) = (1-\pi)\delta_0(\beta_j) + (1-\pi)N(\beta_j|0, \sigma^2\tau^2)$$