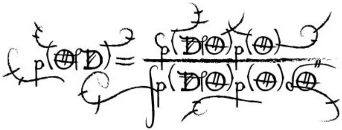


# Bayesian Learning 732A46: Lecture 5


Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank


April 2016

# Lecture overview

- The normal model with a conjugate prior for both $\theta, \sigma^2$.
- Bayesian treatment of the standard linear regression model.
- 'non-informative' prior + Conjugate prior for the linear model.
- Shrinkage (regularization/smoothing) through prior distributions. Connection to the frequentist approach.
- Prediction in the Bayesian linear regression model.

# Normal model - conjugate prior for both $\theta$ and $\sigma^2$

- **Model**

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

- **Conjugate prior**

$$\theta|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$
$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

- Possible to derive $p(\theta, \sigma^2|y)$ by
  1. Using invaluable techniques #1-#3 from **Lecture 3**.
  2. **Ignoring** normalizing constants!
  3. Having **A LOT** of patience.

# Normal model - conjugate prior for both $\theta$ and $\sigma^2$, cont.

- **Posterior**

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

where

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

# Normal model - conjugate prior for both $\theta$ and $\sigma^2$, cont.

- **Posterior**

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

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**

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

- ... or just simulate (marginalization by simulation).

# The standard linear regression model

- The model is

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

  where $i = 1, \ldots, n$. Usually $x_{i1} = 1$ for all $i$ [$\beta_1$ is the intercept].
- **Parameters** $\theta = (\beta_1, \ldots, \beta_k, \sigma^2)'$. **Covariates** $x_i = (1, x_{i2}, \ldots, x_{ik})'$
- **Assumptions**
  1. $E[y_i|x_i, \theta] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ [linear].
  2. $V[y_i|x_i, \theta] = \sigma^2$ [homoscedasticity].
  3. $y_i|x_i, \theta$ conditionally independent for $i = 1, \ldots, n$.
  4. $\varepsilon_i$ are Normal.
- **The notation of**, the posterior distribution

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

  omits explicit conditioning.
- **We are implicitly conditioning** on the $x$'s (covariates) since they are **non-random**.

# The standard linear regression model in matrix form

- The model in **matrix form**

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

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

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

- **The likelihood:** $\left[\Sigma = \sigma^2 \underset{(n\times n)}{I}\right]$

$$p(y|\beta,\sigma^2) \propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta)\right)$$

- $|\Sigma|^{-1/2} = (\sigma^{2n})^{-1/2} = \sigma^{-n}$ and $\Sigma^{-1} = \frac{1}{\sigma^2} I$

# Normal regression - 'non-informative' prior

- The **standard 'non-informative'** prior $p(\beta, \log(\sigma^2)) \propto c$.
- **Equivalently**: $p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$.
- **Summary**: The posterior is $\mathcal{N}\text{-Inv-}\chi^2(\beta_n, \Sigma_n; \nu_n, s_n^2)$

$$\begin{array}{ll} \beta_n = (X'X)^{-1}X'y & \nu_n = n-k \\ \Sigma_n = \sigma^2(X'X)^{-1} & s_n^2 = \frac{1}{n-k}\left(y - X\beta_n\right)'\left(y - X\beta_n\right) \end{array}$$

- **Simulate** from the joint posterior $p(\beta, \sigma^2|y)$:
  1. $\sigma^2|y \sim \text{Inv-}\chi^2(\nu_n, s_n^2)$
  2. $\beta|\sigma^2, y \sim \mathcal{N}(\beta_n, \Sigma_n)$.

# Normal regression - 'non-informative' prior, cont.

- **Some remarks**
  1. $\beta_n$ is the **MLE** of $\beta$ in classical statistic [and $s_n^2$ the **MLE** of $\sigma^2$].
  2. We can show that

$$p(\beta|y) = \int p(\beta|\sigma^2, y)p(\sigma^2|y)d\sigma^2 = t_{n-k}(\beta_n, s_n^2(X'X)^{-1})$$

  $s_n^2(X'X)^{-1}$ - **standard errors for the MLE** of $\beta$.

- A Bayesian analysis with a *'non-informative prior'* gives the same **point estimates** as a Frequentist analysis...
- ... but the Bayesian is **richer** - **knows the whole probability distribution**.
- I have added slides with the derivation. But I will spare you the pain here.

# If you dare, do it at home. Nothing but invaluable techniques #1-#3 (and patience!)

- **Factorize** the posterior (#1)

$$p(\beta, \sigma^2 | y) = p(\beta | \sigma^2, y) p(\sigma^2 | y)$$

- **Determine first** $p(\beta|\sigma^2, y)$ (#2). Use Bayes' theorem and **treat everything but** $\beta$ **as proportionality constants**.
- $p(\beta|\sigma^2, y) = \frac{p(\beta,\sigma^2,y)}{p(\sigma^2,y)} \propto p(y|\beta,\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$
- A **quadratic form** in $\beta$. Thus $\beta|\sigma^2, y \sim \mathcal{N}(\beta_n =?, \Sigma_n =?)$

$$\begin{aligned}
p(\beta|\sigma^2, y) &\propto \exp\left(-\frac{1}{2}(\beta - \beta_n)'\Sigma_n^{-1}(\beta - \beta_n)\right) \\
&= \exp\left(-\frac{1}{2}\left(\beta'\Sigma_n^{-1}\beta - 2\beta'\Sigma_n^{-1}\beta_n + \beta_n'\Sigma_n^{-1}\beta_n\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\beta'\Sigma_n^{-1}\beta - 2\beta'\Sigma_n^{-1}\beta_n\right)\right)
\end{aligned}$$

- **Expand** $\exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$ and **match terms**.

# Derivations, cont.

- **Expanding**

$$\begin{aligned}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right) &= \exp\left(-\frac{1}{2\sigma^2}(y'y-2\beta'X'y+\beta'X'X\beta)\right)\\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\beta'X'X\beta-2\beta'X'y)\right)\end{aligned}$$

- **Match** $\Sigma_n$:

$$\beta'X'X\beta/\sigma^2 = \beta'\Sigma_n^{-1}\beta \implies \Sigma_n^{-1} = \frac{1}{\sigma^2}X'X \implies \Sigma_n = \sigma^2(X'X)^{-1}$$

- For $\beta_n$, first rewrite

$$2\beta'X'y/\sigma^2 = 2\beta'\underbrace{\Sigma_n^{-1}\Sigma_n}_{I}X'y/\sigma^2$$

- **Match** $\beta_n$:

$$2\beta'\Sigma_n^{-1}\Sigma_nX'y/\sigma^2 = 2\beta'\Sigma_n^{-1}\beta_n \implies \beta_n = (X'X)^{-1}X'y$$

- **We conclude** that $p(\beta|\sigma^2, y) = \mathcal{N}(\beta_n, \Sigma_n)$

$$\begin{aligned}\beta_n &= (X'X)^{-1}X'y \\ \Sigma_n &= \sigma^2(X'X)^{-1}.\end{aligned}$$

- Derive $p(\sigma^2|y)$ (#3)

$$p(\sigma^2|y) = \frac{p(\beta, \sigma^2|y)}{p(\beta|\sigma^2, y)} \propto \frac{p(y|\beta, \sigma^2)p(\sigma^2)}{p(\beta|\sigma^2, y)}$$

- **Standard trick**: LHS does not depend on $\beta$ ($\beta$ cancels on RHS).
- Evaluate RHS using $\beta = \beta_n$ (simplifies the denominator)

$$\begin{aligned}\frac{p(y|\beta_n, \sigma^2)p(\sigma^2)}{p(\beta_n|\sigma^2, y)} &\propto \frac{\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta_n)'(y - X\beta_n)\right)\sigma^{-2}}{|\Sigma_n|^{-1/2} \exp\left(-\frac{1}{2}(\beta_n - \beta_n)'\Sigma_n^{-1}(\beta_n - \beta_n)\right)} \\ &= \frac{\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta_n)'(y - X\beta_n)\right)\sigma^{-2}}{|\Sigma_n|^{-1/2}}\end{aligned}$$

- $|\Sigma_n| = \left|\sigma^2(X'X)^{-1}\right| = \left|\sigma^2 I(X'X)^{-1}\right| = \overbrace{\left|\sigma^2 I\right|}^{=\sigma^{2k}}\left|(X'X)^{-1}\right|.$
- Thus $|\Sigma_n|^{-1/2} = \left(\sigma^{2k}\right)^{-1/2}\left(\left|(X'X)^{-1}\right|\right)^{-1/2} \propto \sigma^{-k}$, and

$$p(\sigma^2|y) \propto \sigma^{-n+k-2}\exp\left(-\frac{1}{2\sigma^2}(y - X\beta_n)'(y - X\beta_n)\right).$$

- $p(\sigma^2|y) = \text{Inv-}\chi^2(\nu_n, s_n^2)$ if it is **proportional to**

$$\sigma^{-2(\nu_n/2+1)}\exp\left(-\frac{\nu_n s_n^2}{2\sigma^2}\right).$$

- **Rewrite** and **match terms**

$$p(\sigma^2|y) \propto \sigma^{-2(\overbrace{(n-k)}^{\nu_n}/2+1)}\exp\left(-\frac{\overbrace{n-k}^{\nu_n}}{2\sigma^2}\underbrace{\frac{1}{(n-k)}(y - X\beta_n)'(y - X\beta_n)}_{s_n^2}\right)$$

# Derivations, **final slide**!

- **We have proven** that the posterior is $\mathcal{N}$-Inv-$\chi^2(\beta_n, \Sigma_n; \nu_n, s_n^2)$

$$\begin{array}{ll}\beta_n = (X'X)^{-1}X'y & \nu_n = n-k \\ \Sigma_n = \sigma^2(X'X)^{-1} & s_n^2 = \frac{1}{n-k}\left(y - X\beta_n\right)'\left(y - X\beta_n\right)\end{array}$$

# Normal regression - Conjugate prior for $\beta, \sigma^2$

- An informative prior is helpful for **model regularization**.
- **Conjugate prior** $p(\beta, \sigma^2) = \mathcal{N}\text{-Inv-}\chi^2(\beta_0, \sigma^2\Omega_0^{-1}; \nu_0, s_0^2)$,

$$
\begin{aligned}
p(\beta|\sigma^2) &= \mathcal{N}(\beta_0, \sigma^2\Omega_0^{-1}) \\
p(\sigma^2) &= \text{Inv-}\chi^2(\nu_0, s_0^2)
\end{aligned}
$$

- The role of the **hyperparameters** in the prior.
  (1) $\beta_0$ - the mean. $\beta_0 = 0$ common choice.
  (2) $\Omega_0$ - the "precision". Common choice $\Omega_0 = \lambda I$ $[\Omega_0^{-1} = \frac{1}{\lambda} I]$.
  **Larger** $\lambda \implies$ **prior concentrates more around** $\beta_0$. We can even have

$$\Omega_0 = \text{diag}(\lambda_1, \ldots, \lambda_k).$$

  (3) $\nu_0$ - prior degrees of freedom.
  (4) $s_0^2$ - prior average sum of squares.
- With $\beta_0 = 0$, (2) is an example of **model regularization through the prior**. Tackles **over-fitting problems** that occur in models with many parameters.

# The posterior with the conjugate prior for $\beta, \sigma^2$

- **Tedious** algebra with 'non-informative' prior.
- A **nightmare** here - **but still only** invaluable techniques #1-#3.
- **It's form:** a matrix version of the normal model with a conjugate on $\theta, \sigma^2$.
- Let $\hat{\beta} = (X'X)^{-1}X'y$ [**'Non-informative'** case, **MLE**].
- **The posterior** $p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y)$

$$
\begin{aligned}
p(\beta|\sigma^2, y) &= \mathcal{N}(\beta_n, \sigma^2\Omega_n^{-1}) \\
p(\sigma^2|y) &= \text{Inv-}\chi^2(\nu_n, s_n^2)
\end{aligned}
$$

$$
\begin{array}{ll}
\beta_n = (X'X + \Omega_0)^{-1}(X'X\hat{\beta} + \Omega_0\beta_0) & \Omega_n = X'X + \Omega_0 \\
\nu_n = \nu_0 + n & \nu_n s_n^2 = \nu_0 s_0^2 + y'y + \beta_0'\Omega_0\beta_0 - \beta_n'\Omega_n\beta_n
\end{array}
$$

- **Note**: the posterior mean is a weighted average of data and prior information.

# Shrinkage illustration with the conjugate prior for $\beta, \sigma^2$

- **Shrinkage illustration**:

  With $X'X = I$ and $p(\beta|\sigma^2) = \mathcal{N}(\beta_0, \sigma^2/\lambda I)$ $[\Omega_0 = \lambda I]$

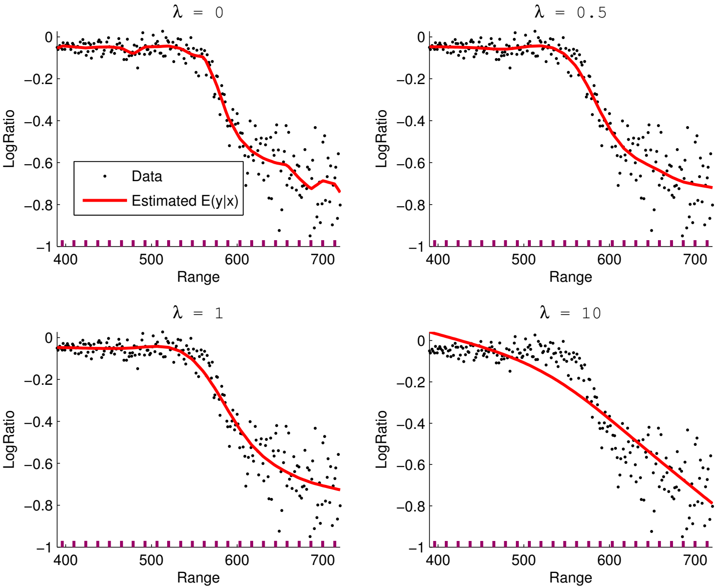
$$\beta_n = (I + \lambda I)^{-1}(I\hat{\beta} + \lambda I \beta_0) = \overbrace{\frac{1}{1+\lambda}}^{0<w_1<1}\hat{\beta} + \overbrace{\frac{\lambda}{1+\lambda}}^{0<w_2<1}\beta_0 \quad [w_1 + w_2 = 1].$$

- **Note that**
  1. $\lambda \to 0 \implies \beta_n \to \hat{\beta}$.
  2. $\lambda \to \infty \implies \beta_n \to \beta_0$.

- **'Non-informative' priors** $[\lambda = 0]$ **do not shrink**!...

- ... **Neither does Frequentist OLS nor MLE estimates!**

# Example: Bayesian spline with shrinkage prior

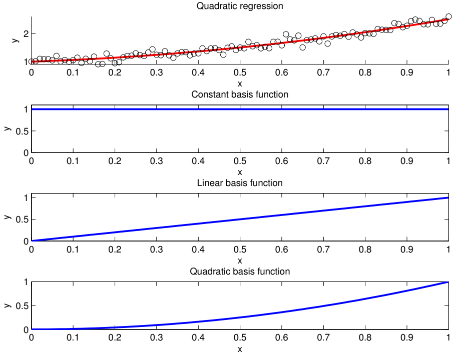


# Polynomial regression as a simple linear regression

- Consider only a **single covariate** for simplicity.
- A general regression model with **additive noise**

$$y_i = f(x_i; \beta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- **Polynomial regression**
  - $f(x_i; \beta) = \beta_0 + \beta_1 x_i$ - linear regression
  - $f(x_i; \beta) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ - quadratic regression
  - $f(x_i; \beta) = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p$ - polynomial order $p$.



# Polynomial regression as a simple linear regression, cont.

- **Can be written**

$$y = X_p\beta + \varepsilon, \quad X_p = (1, x, x^2, \ldots x^p) \in n \times (p+1).$$

- $x$ is **basis expanded** $X_p = (h_0(x), h_1(x), h_2(x), \ldots h_p(x))$ where

$$h_j(x) = x^j, \quad j = 1, \ldots, p, \text{ are the basis functions.}$$

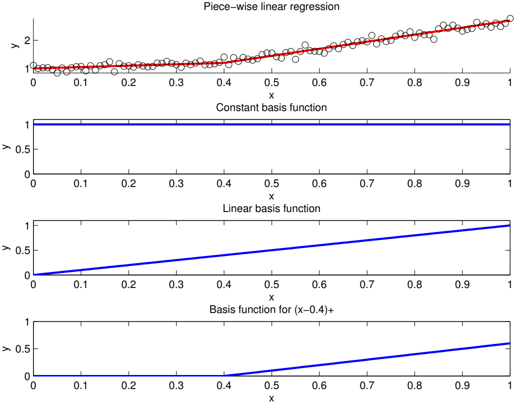
- **Note:** The model is **Non-linear** in data but **still linear in parameter**
- Estimation as before but with $X_p$ in place of $X$.
- **Problem:** Polynomials are **too global** - fit becomes unstable.

# Shrinkage in a spline regression model

- **Splines** to the rescue! Like polynomials but **more local**.
- **Example:** Truncated *power splines*. Basis functions

$$h_j(x) = \begin{cases} (x - \xi_j)^a & \text{if } x > \xi_j \\ 0 & \text{otherwise.} \end{cases},$$

where $\xi_j$ $(j = 1, \ldots, p)$ are the **knots** which are given.



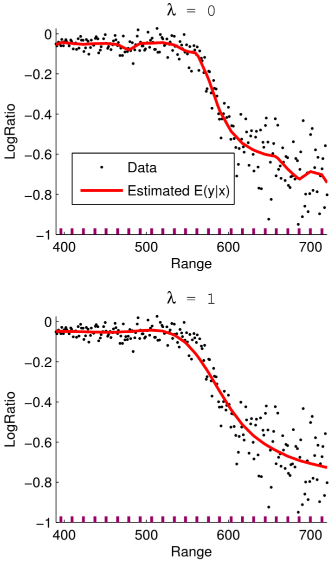
# Shrinkage in a spline regression model

- **Note**: given the knots, the **spline regression model** is a linear regression of $y$ on the basis expanded matrix

$$X_p = (1, x, h_1(x), \ldots h_p(x))$$

  (common to include **an intercept** + **a linear basis** function).
- Estimation as in the linear regression. Just change $X$ for $X_p$.
- Typically many knots. **Regularization** required for a smooth fit.
- **Let's see the figure again**!

# Shrinkage in Bayesian spline (note the knots!)



- Frequentists instead shrink by minimizing a **penalized** RSS.
- Residual Sum of Squares (RSS):

$$\mathrm{RSS}(\beta) = (y - f(X;\beta))'(y - f(X;\beta)).$$

- Example **ridge regression**:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} RSS(\beta) + \lambda\beta'\beta \longrightarrow \hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y$$

- The same as the **Bayesian posterior mean** $\beta_n$ we have derived with $\beta_0 = 0$ and $\Omega_0 = \lambda I$...

# Shrinkage: Frequentist vs Bayesian

- Frequentists instead shrink by minimizing a **penalized** RSS.
- Residual Sum of Squares (RSS):

$$\mathrm{RSS}(\beta) = (y - f(X;\beta))'(y - f(X;\beta)).$$

- Example **ridge regression**:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} RSS(\beta) + \lambda\beta'\beta \longrightarrow \hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y$$

- The same as the **Bayesian posterior mean** $\beta_n$ we have derived with $\beta_0 = 0$ and $\Omega_0 = \lambda I$...
- ... the frequentists are indeed using "prior information"... **but they are hiding it**!
- The Bayesian shrinkage prior is **interpretable**. Nothing ad hoc.

# Other Shrinkage priors

- Other shrinkage priors can be used but they are **not conjugate**.
- MCMC methods can be used for estimation.
- The famous frequentist **Lasso variable selection** method is equivalent to **the posterior mode** using the prior

$$p(\beta_k|\sigma^2) = \frac{\lambda}{2\sigma^2} \exp\left(-\lambda \frac{|\beta_k|}{\sigma^2}\right) \quad \left[\mathrm{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right)\right].$$

- Laplace distribution - **heavy tails**.
- **Laplace prior**: many $\beta_k$ are close to zero, but some $\beta_i$ may be very large.
- Normal distribution - **light tails**.
- **Normal prior**: most $\beta_k$ are fairly equal in size, and no single $\beta_k$ can be very much larger than the other ones.

# Estimating the shrinkage parameter $\lambda$ from the data

- So far implicitly assumed that $\lambda$ **is known**.
- Crossvalidation is one possibility. But this **is not** a Bayesian solution.
- **Question**: What would a Bayesian solution be?

# Estimating the shrinkage parameter $\lambda$ from the data

- So far implicitly assumed that $\lambda$ **is known**.
- Crossvalidation is one possibility. But this **is not** a Bayesian solution.
- **Question**: What would a Bayesian solution be?
- **Clue**: Bayesians treat *any* **unknown quantity** as...

# Estimating the shrinkage parameter $\lambda$ from the data

- So far implicitly assumed that $\lambda$ **is known**.
- Crossvalidation is one possibility. But this **is not** a Bayesian solution.
- **Question**: What would a Bayesian solution be?
- **Clue**: Bayesians treat *any* **unknown quantity** as...
- ... a **random variable**!
- Treat $\lambda$ as a r.v. Inference through $p(\beta, \sigma^2, \lambda | y)$
- Assign a prior $p(\lambda)$ and derive $p(\lambda | y)$.

- The **joint posterior** factorizes (#1)

$$p(\beta, \sigma^2, \lambda | y) = p(\beta | \sigma^2, \lambda, y) p(\sigma^2 | \lambda, y) p(\lambda | y),$$

where (#1-#2 to derive)

$$\begin{array}{rcl}
\textbf{Prior} & \rightarrow & \textbf{Posterior} \\
\beta | \sigma^2, \lambda \sim \mathcal{N}(0, \sigma^2 \Omega_0^{-1}) & \rightarrow & \beta | \sigma^2, \lambda, y \sim \mathcal{N}(\beta_n, \sigma^2 \Omega_n^{-1}) \\
\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, s_0^2) & \rightarrow & \sigma^2 | \lambda, y \sim \text{Inv-}\chi^2(\nu_n, s_n^2) \\
\lambda \sim p(\lambda) & \rightarrow & \lambda | y \sim \sqrt{\frac{|\Omega_0|}{|\Omega_n|}} \left( \frac{\nu_n s_n^2}{2} \right)^{-\nu_n/2} p(\lambda)
\end{array}$$

and

$$\begin{array}{ll}
\beta_n = (X'X + \Omega_0)^{-1} X'y & \Omega_n = X'X + \Omega_0 \\
\nu_n = \nu_0 + n & \nu_n s_n^2 = \nu_0 s_0^2 + y'y - \beta_n' \Omega_n \beta_n
\end{array}$$

# Prediction in linear regression

- **Predict the outcome** $\tilde{y}$ for a set of observations with covariates $\tilde{X}$
- **Posterior predictive density** [implicitly conditioning on $X$ and $\tilde{X}$]

$$\begin{aligned} p(\tilde{y}|y) &= \int_{\sigma^2}\int_{\beta} p(\tilde{y}|y,\beta,\sigma^2)p(\beta,\sigma^2|y)d\beta d\sigma^2 \\ &= \int\int p(\tilde{y}|\beta,\sigma^2)p(\beta,\sigma^2|y)d\beta d\sigma^2 \end{aligned}$$

  if $\tilde{y}$ and $y$ are conditionally independent (given $\beta, \sigma^2$).

- **We can simulate** from $p(\tilde{y}|y)$ by
  1. $\beta, \sigma^2 \sim p(\beta, \sigma^2|y)$
  2. $\tilde{y}|\beta, \sigma^2 \sim p(\tilde{y}|\beta, \sigma^2) = \mathcal{N}(\tilde{X}\beta, \sigma^2 I)$.
- Step 1.: First $\sigma^2$ and then $\beta|\sigma^2$ $\left[p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y)\right]$.
- If the **shrinkage** $\lambda$ **is estimated** use $p(\beta, \sigma^2, \lambda|y)$ instead.