SL = BAYESIAN LINEAR MODELS

We revisit linear regression & logistic regression from a Bayesian point of view. For background information on Bayesian statistics, see MS: BAYESIAN STATISTICS .

I. BAYESIAN LINEAR REGRESSION

Consider a learning sample $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where (X_i, Y_i) are iid, and are assumed to arise from a linear model $Y = X\beta + \varepsilon$, where

$$\underset{(n\times 1)}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \underset{n\times(d+1)}{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix}, \quad \underset{(n\times 1)}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_d \end{pmatrix}$$

see pages 1/2 in SL: LINEAR REGRESSION .

→ The frequentist approach to linear regression assumes that the vector of parameters β is fixed and unknown. It is estimated by maximizing the likelihood function (≡ least squares estimate under normal errors ε).

→ The Bayesian treatment considers the vector β to be random, with prior distribution $f(\beta)$:

$$\boxed{\begin{aligned} &Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\varepsilon \mid 0, \gamma^{-1} I_n) \\ &\qquad + \text{ prior } \quad \beta \sim \mathcal{N}(\beta \mid 0, \alpha^{-1} I_{d}) = f(\beta) \end{aligned}}$$

BAYESIAN LINEAR MODEL

For notational convenience, we assume that $\beta \in \mathbb{R}^d$,

Remarks : (i) More generally, we can assume that $f(\beta) = \mathcal{N}(\beta \mid m_0, S_0)$. Subsequent calculations can be easily adapted.

(ii) We consider two cases :

- β unknown , γ known (sections I.1 & I.2)
- β unknown , γ unknown (section I.3)

In addition, we consider the case where hyperpriors are introduced for α and γ ; the so-called Empirical Bayes / Type II ML / Evidence approximation approach (section I.4).

In each case, we are interested in

- the posterior distribution of β, having observed $\mathcal{L}_n$ (→ useful for the construction of credible intervals)
- the predictive distribution of Y given $\mathcal{L}_n$, and a new input point x .

I.1. Posterior distribution (γ known).

- The posterior distribution is $\underbrace{f(\beta \mid \mathcal{L}_n)}_{\text{posterior}} \underset{\uparrow \text{ "proportional to"}}{\propto} \underbrace{f(\mathcal{L}_n \mid \beta)}_{\text{likelihood}} \underbrace{f(\beta)}_{\text{prior}}$,

where

- $f(\mathcal{L}_n \mid \beta) = \left(\dfrac{\gamma}{2\pi}\right)^{n/2} \exp\left\{-\dfrac{\gamma}{2}(y - X\beta)^t (y - X\beta)\right\}$,

$f(\beta) = \left(\dfrac{\alpha}{2\pi}\right)^{1/2} \exp\left\{-\dfrac{\alpha}{2}\beta^t \beta\right\}$.

The product of the likelihood by the prior is proportional to:

$$\sim \exp\left\{-\frac{1}{2}\left[\underbrace{\gamma\beta^t X^t X\beta - 2\gamma y^t X\beta + \gamma y^t y}_{\text{from the likelihood}} + \underbrace{\alpha\beta^t\beta}_{\text{from the prior}}\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\beta^t(\alpha I_d + X^t X)\beta - (\gamma X^t y)^t\beta + \underset{\text{indpt of } \beta}{\text{constant}}\right]\right\}$$

$\uparrow$

We recognize here the expression of the multivariate normal density.

$\Rightarrow f(\beta \mid \mathcal{L}_n) = \mathcal{N}(\beta \mid m_n, S_n)$.

To find the expression of m_n and S_n, compare the terms in the expression above with

$$(\beta - m_n)^t S_n^{-1}(\beta - m_n) = \beta^t S_n \beta - 2 m_n^t S_n^{-1}\beta + m_n^t S_n^{-1} m_n$$

We immediately get: $\begin{cases} S_n^{-1} = \alpha I_d + \gamma X^t X \\ m_n = \gamma S_n X^t y \end{cases}$.

(*)

Summary: $X = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(\varepsilon \mid 0, \gamma^{-1} I_n)$
$\beta \sim \mathcal{N}(\beta \mid 0, \alpha^{-1} I_d)$

Posterior is $f(\beta \mid \mathcal{L}_n) = \mathcal{N}(\beta \mid m_n, S_n)$, with

$$\begin{cases} m_n = \gamma S_n X^t y \\ S_n^{-1} = \alpha I_d + \gamma X^t X \end{cases}$$

× Remarks: (i) Assuming more generally that $f(\beta) = \mathcal{N}(\beta \mid m_0, S_0)$, we easily derive the expression for the posterior mean & covariance:

$m_n = S_n(S_0^{-1} m_0 + \gamma X^t y)$ and $S_n^{-1} = S_0^{-1} + \gamma X^t X$.

(ii) Bayesian linear regression & Ridge Regression (RR)

The log of the posterior distribution is:

$$\log f(\beta \mid \mathcal{L}_n) = -\frac{\gamma}{2}\sum_{i=1}^{n}(y_i - x_i^t\beta)^2 - \frac{\alpha}{2}\beta^t\beta$$

$$= -\frac{\gamma}{2}\Big\{\underbrace{\sum_{i=1}^{n}(y_i - x_i^t\beta)^2 + \frac{\alpha}{\gamma}\beta^t\beta}_{\text{RSS}_2(\alpha/\gamma)}\Big\}$$

$\uparrow$ see p.5 in SL: RR AND LASSO.

- Consequences:

→ α/γ is a tuning parameter, and quantifies the trade-off between the goodness-of-fit term ($\equiv$ likelihood) and the penalty ($\equiv$ prior). It can be estimated using cross-validation techniques.

→ Ridge solution = MAP estimator.

(iii) Points arriving sequentially.

Assuming a stream of observations

$$(x_1, y_1) \to (x_2, y_2) \to \cdots \to (x_n, y_n) \to (x_{n+1}, y_{n+1}) \to \cdots;$$

the posterior distribution after n points are collected is $\mathcal{N}(\beta \mid m_n, S_n)$.

A new observation (x_{n+1}, y_{n+1}) has density / likelihood

$$f(y_{n+1} \mid x_{n+1}, \beta) = \left(\frac{\gamma}{2\pi}\right)^{1/2}\exp\left[-\frac{\gamma}{2}(y_{n+1} - \beta^t x_{n+1})^2\right],$$

since $Y_{n+1} = \beta^t x_{n+1} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \gamma^{-1})$.

The posterior distribution, having observed $(n+1)$ points, is proportional (w.r.t. β) to:

$$\exp\left\{-\frac{1}{2}\left[(\beta-m_n)^t S_n^{-1}(\beta-m_n) + \gamma(y_{n+1}-\beta^t x_{n+1})^2\right]\right\}$$

$$\shortparallel$$

$$\left[\beta^t(S_n^{-1}+\gamma x_{n+1}x_{n+1}^t)\beta \right.$$
$$-2\beta^t(S_n^{-1}m_n + \gamma x_{n+1}y_{n+1})$$
$$\left. + \text{constant indpt of } \beta \right]$$

$\uparrow$

Compare this expression with $(\beta-m_{n+1})^t S_{n+1}^{-1}(\beta-m_{n+1})$, appearing in the posterior $f(\beta \mid \mathcal{L}_{n+1}) = \mathcal{N}(\beta \mid m_{n+1}, S_{n+1})$.

We see that
$$\begin{cases} S_{n+1}^{-1} = S_n^{-1} + \gamma x_{n+1} x_{n+1}^t \\ m_{n+1} = S_{n+1}(S_n^{-1}m_n + \gamma x_{n+1}y_{n+1}) \end{cases},$$

with $m_0 = 0$ and $S_0 = \alpha^{-1}I$.

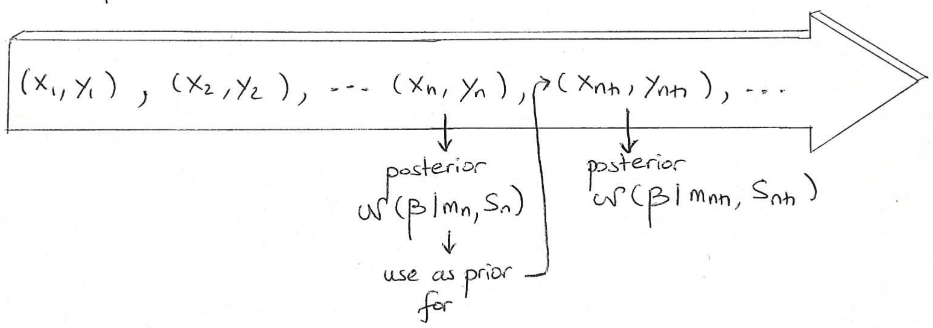
$\uparrow$ coincides with the formula (*) on page 3 since:

- $$S_{n+1}^{-1} = S_{n-1}^{-1} + \gamma x_{n+1}x_{n+1}^t + \gamma x_n x_n^t$$
$$\vdots$$
$$= S_0^{-1} + \gamma \sum_{i=1}^{n+1} x_i x_i^t$$
$$= \alpha I + \gamma X^t X$$

- $$m_{n+1} = S_{n+1}\Big(S_n^{-1}\big[\underbrace{S_n(S_{n-1}^{-1}m_{n-1}+\gamma x_n y_n)}_{\overset{\shortparallel}{m_n}}\big] + \gamma x_{n+1}y_{n+1}$$
$$= S_{n+1}(S_{n-1}^{-1}m_{n-1} + \gamma x_n y_n + \gamma x_{n+1}y_{n+1})$$
$$\vdots$$
$$= S_{n+1}(\cancel{S_0^{-1}m_0}^{\,0} + \gamma \sum_{i=1}^{n+1} x_i y_i)$$
$$= \gamma S_{n+1} X^t y$$

Csq: in a sequatial setting,

posterior distribution $f(\beta \mid \mathcal{L}_n)$ = prior distribution on β for a new observation (x_{n+1}, y_{n+1}).


(x_1,y_1), (x_2,y_2), ... (x_n,y_n), (x_{n+1},y_{n+1}), ...
posterior
$\mathcal{N}(\beta \mid m_n, S_n)$
use as prior for
posterior
$\mathcal{N}(\beta \mid m_{n+1}, S_{n+1})$

I.2. Predictive distribution (γ known).

- Given a new input point x, the predictive distribution is

$$f(y \mid \mathcal{L}_n, x) = \int f(y, \beta \mid \mathcal{L}_n, x, \alpha, \gamma)\, d\beta$$
$$= \int \underbrace{f(y \mid \beta, x, \gamma)}_{\substack{\uparrow \\ \text{target var. distrib} \\ \mathcal{N}(y \mid x^t\beta,\, \gamma^{-1})}} \underbrace{f(\beta \mid \mathcal{L}_n, \alpha, \gamma)}_{\substack{\uparrow \\ \text{posterior distrib} \\ \mathcal{N}(\beta \mid m_n, S_n)}} \underbrace{d\beta}_{\substack{\uparrow \\ \beta \text{ is} \\ \text{integrated} \\ \text{out}}}$$

= convolution of two Gaussian distributions

$\Rightarrow$ still Gaussian.

$$f(y \mid \mathcal{L}_n, x) = \mathcal{N}(y \mid m_n^t x,\ \underbrace{\gamma^{-1} + x^t S_n x}_{=:\ \sigma_n^2(x)})$$

$\uparrow$

More generally, if $p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$
$p(y|x) = \mathcal{N}(y \mid Ax+b, L^{-1})$,

then $p(y) = \mathcal{N}(y \mid A\mu+b, L^{-1}+A\Lambda^{-1}A^t)$, see Bishop p. 93

Remarks: (i) As $n \to +\infty$, $\sigma_n^2(x) \longrightarrow \gamma^{-1}$ ($\equiv$ noise variance)

Indeed,

$\sigma_{n+1}^2(x) = \gamma^{-1} + x^t S_{n+1} x$, where

$$S_{n+1} = \left(S_n^{-1} + \gamma\, x_{n+1} x_{n+1}^t \right)^{-1} \quad \text{(page 5)}$$

$$= S_n - \frac{(S_n x_{n+1} \gamma^{1/2})(\gamma^{1/2} x_{n+1}^t S_n)}{1 + \gamma x_{n+1}^t S_n x_{n+1}}$$

$$= S_n - \gamma \frac{S_n x_{n+1} x_{n+1}^t S_n}{1 + \gamma x_{n+1}^t S_n x_{n+1}}$$

Toolbox:

$$(A + vv^t)^{-1} \overset{\shortparallel}{} A^{-1} - \frac{(A^{-1}v)(v^t A^{-1})}{1 + v^t A^{-1} v}$$

$$\Downarrow$$

$$\sigma_{n+1}^2(x) = \gamma^{-1} + x^t \left(\text{—}\,''\,\text{—} \right) x$$

$$= \sigma_n^2(x) - \underbrace{\gamma \frac{x^t \overset{\text{PSD}}{\left(S_n x_{n+1} x_{n+1}^t S_n\right)} x}{1 + \gamma \underbrace{x_{n+1}^t S_n x_{n+1}}_{\substack{\text{non-negative} \\ \text{since } S_n \text{ is} \\ \text{positive semi-def.} \\ \text{(PSD)}}}}}_{\geqslant 0}$$

Thus, $\sigma_{n+1}^2(x) \leqslant \sigma_n^2(x)$, as required. ■

(ii) In a Bayesian linear regression setting, the posterior can be computed analytically. Alternatively, we could sample points $\beta_i \sim \mathcal{N}(m_n, S_n)$ from the posterior distribution, and consider the Monte-Carlo approximation to the predictive distribution:

$$\frac{1}{M} \sum_{i=1}^{M} f(y \mid \beta_i, x, \gamma).$$

Will be useful in more complex settings.

I.3. Case when β and γ are unknown.

We assume now that $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \gamma^{-1} I_n)$, with γ unknown.

↳ the conjugate prior on (β, γ) is the normal-gamma distribution (see MS: BAYESIAN STATISTICS):

$$f(\beta, \gamma) = \mathcal{N}(\beta \mid m_0, \gamma^{-1} S_0)\, \text{Gamma}(\gamma \mid a_0, b_0),$$

where

$$\text{Gamma}(\gamma \mid a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \gamma^{a_0 - 1} e^{-b_0 \gamma}, \quad \gamma > 0$$

↑ see p. 21 in PT: POPULAR DISTRIBUTIONS

↳ the posterior distribution is $f(\beta, \gamma \mid \mathcal{L}_n) \propto f(\mathcal{L}_n \mid \beta, \gamma) f(\beta, \gamma)$

In the log-space,

$$= \log f(\beta, \gamma) + \log f(\mathcal{L}_n \mid \beta, \gamma)$$

$$= \boxed{\log \mathcal{N}(\beta \mid m_0, \gamma^{-1} S_0)} + \boxed{\log \text{Gamma}(\gamma \mid a_0, b_0)}$$
$$+ \boxed{\log \mathcal{N}(y \mid X\beta, \gamma^{-1} I_n)}$$

$$= \boxed{\frac{d}{2} \log \gamma - \frac{1}{2} \log |S_0| - \frac{\gamma}{2} (\beta - m_0)^t S_0^{-1} (\beta - m_0)}$$
$$\boxed{- b_0 \gamma + (a_0 - 1) \log \gamma}$$
$$\boxed{+ \frac{n}{2} \log \gamma - \frac{\gamma}{2} \sum_{i=1}^{n} (y_i - \beta^t x_i)^2}$$

+ constant term in β, γ.

We selected the prior distribution such that the posterior belongs to the same family of distribution.

$\Rightarrow f(\beta, \gamma \mid \mathcal{L}_n) =$ normal-gamma.

To find the parameters of the normal-gamma distribution, we write the posterior as a product of two densities (one will correspond to the normal term, the other to the gamma term):

$$f(\beta, \gamma \mid \mathcal{L}_n) = \underbrace{f(\beta \mid \mathcal{L}_n, \gamma)}_{\uparrow} f(\gamma \mid \mathcal{L}_n).$$

We first identify this term, and collect in the expression at the bottom of page 8 all terms involving β:

$$-\frac{\gamma}{2} \beta^t (X^t X + S_0^{-1}) \beta + \gamma \beta^t (S_0^{-1} m_0 + X^t y) + \text{cst in } \beta$$

↑ Compare the terms with those appearing in $f(\beta \mid \mathcal{L}_n, \gamma) = \mathcal{N}(\beta \mid m_n, \gamma^{-1} S_n)$:

$$-\frac{1}{2} (\beta - m_n)^t (\gamma^{-1} S_n)^{-1} (\beta - m_n)$$

$$= -\frac{\gamma}{2} \beta^t S_n^{-1} \beta + \gamma \beta^t S_n^{-1} m_n + \text{cst in } \beta$$

we see that

$$\left(\begin{array}{l} \gamma S_n^{-1} = \gamma (X^t X + S_0^{-1}) \\ \gamma (S_0^{-1} m_0 + X^t y) = \gamma S_n^{-1} m_n . \end{array} \right.$$

$$\Rightarrow \boxed{\begin{array}{l} f(\beta \mid \mathcal{L}_n, \gamma) = \mathcal{N}(\beta \mid m_n, \gamma^{-1} S_n), \text{ with} \\ \qquad m_n = S_n (S_0^{-1} m_0 + X^t y) \\ \qquad S_n^{-1} = S_0^{-1} + X^t X \end{array}}$$

It remains to identify all remaining terms to identify the parameters of the Gamma distribution $\text{Gamma}(\gamma \mid a_n, b_n)$

Note that the term $\left(\frac{d}{2} \log \gamma \right)$ appearing in the expression on the bottom of page 8 is incorporated into the expression of $f(\beta \mid \mathcal{L}_n, \gamma)$.

$$\Rightarrow \log f(\gamma \mid \mathcal{L}_n) = \boxed{-b_0 \gamma + (a_0 - 1) \log \gamma}$$

$$\boxed{+ \frac{n}{2} \log \gamma - \frac{\gamma}{2} \sum_{i=1}^{n} y_i^2}$$

$$\boxed{- \frac{\gamma}{2} m_0^t S_0^{-1} m_0}$$

$$\boxed{+ \frac{\gamma}{2} m_n^t S_n^{-1} m_n} \rightarrow$$

from completing the squares

= log of a gamma distribution, with parameters

$$a_n = a_0 + \frac{n}{2}$$

$$b_n = b_0 + \frac{1}{2} \left(\sum y_i^2 + m_0^t S_0^{-1} m_0 - m_n^t S_n^{-1} m_n \right)$$

Summary:

- prior on (β, γ) is $f(\beta, \gamma) = \mathcal{N}(\beta \mid m_0, \gamma^{-1} S_0) \text{Gamma}(\gamma \mid a_0, b_0)$
- posterior is $f(\beta, \gamma \mid \mathcal{L}_n) = \mathcal{N}(\beta \mid m_n, \gamma^{-1} S_n) \text{Gamma}(\gamma \mid a_n, b_n)$, with
  - → $m_n = S_n (S_0^{-1} m_0 + X^t y)$
  - → $S_n^{-1} = S_0^{-1} + X^t X$
  - → a_n, b_n given above

Before moving on to the predictive distribution, we review a useful result:

"Bayesian view" of Student distribution: Assume $X \sim \mathcal{N}(\mu, s\lambda^{-1})$

$\lambda \sim \text{gamma}(a, b)$.

Then X has a Student's t-distribution.

↳ Indeed,

$$f(x) = \int_0^{+\infty} f(x, \lambda)\, d\lambda$$

$$= \int_0^{+\infty} f(x \mid \lambda) f(\lambda)\, d\lambda$$

$$= \int_0^{+\infty} \left(\frac{\lambda}{2\pi s}\right)^{1/2} e^{-\frac{\lambda}{2s}(x-\mu)^2} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}\, d\lambda$$

$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi s}\right)^{1/2} \underbrace{\int_0^{+\infty} \lambda^{a-\frac{1}{2}} e^{\boxed{-\lambda\left(b+\frac{(x-\mu)^2}{2s}\right)}}\, d\lambda}_{u^{-a+\frac{1}{2}} \int_0^{+\infty} e^{-z} u^{-1} z^{a-\frac{1}{2}}\, dz}$$

Change of variable $z = \lambda u$,
$u = b + \frac{1}{2s}(x-\mu)^2$

$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi s}\right)^{1/2} u^{-a-1/2} \underbrace{\int_0^{+\infty} z^{a-\frac{1}{2}} e^{-z}\, dz}_{=\Gamma(a+\frac{1}{2})}$$

$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi s}\right)^{1/2} \Gamma\left(a+\tfrac{1}{2}\right)\left(b + \frac{(x-\mu)^2}{2s}\right)^{-a-\frac{1}{2}}$$

Put $\begin{cases} k = 2a & \longrightarrow \quad a = k/2 \\ \tau = \frac{a}{b} s^{-1} & \longrightarrow \quad b = \frac{a}{\tau s} = \frac{k}{2\tau s} \end{cases}$

$$f(x) = \left(\frac{k}{2\tau s}\right)^{k/2} \frac{1}{\Gamma(k/2)}\left(\frac{1}{2\pi s}\right)^{1/2} \Gamma\left(\frac{k+1}{2}\right) \underbrace{\left(\frac{k}{2\tau s} + \frac{(x-\mu)^2}{2s}\right)^{-\frac{k+1}{2}}}_{\left(\frac{k}{2\tau s}\right)^{-\frac{k+1}{2}}\left(1 + \frac{\tau}{k}(x-\mu)^2\right)^{-\frac{k+1}{2}}}$$

$$\boxed{f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(k/2)}\left(\frac{\tau}{\pi k}\right)^{1/2}\left(1 + \frac{\tau(x-\mu)^2}{k}\right)^{-\frac{k+1}{2}}}$$

$k = 2a$
$\tau = \frac{a}{b} s^{-1}$

↑ Compare with the expression page 39 in PT: POPULAR DISTRIBUTIONS

$$\boxed{X \sim t(k, \mu, \tau)}$$

Remark: $\left(1 + \frac{\tau(x-\mu)^2}{k}\right)^{-\frac{k+1}{2}} = \exp\left\{-\frac{k+1}{2}\log\left(1 + \frac{\tau(x-\mu)^2}{k}\right)\right\}$

$$\approx \exp\left\{-\frac{k+1}{2}\left(\frac{\tau(x-\mu)^2}{k} + O(k^{-2})\right)\right\}$$

$$\longrightarrow \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right) \text{ as } k \to +\infty$$

so that

$$X \xrightarrow{d} \mathcal{N}(\mu, \tau^{-1}). \quad ∎$$

- Back to the predictive distribution.

$$f(y \mid \mathcal{D}_n, x) = \iint f(y \mid \beta, \gamma, x)\, f(\beta, \gamma \mid \mathcal{D}_n)\, d\beta\, d\gamma$$

$$= \iint \mathcal{N}(y \mid x^t\beta, \gamma^{-1})\, \mathcal{N}(\beta \mid m_n, \gamma^{-1}S_n)\, \text{Gamma}(\gamma \mid a_n, b_n)\, d\beta\, d\gamma$$

x First, compute the integral w.r.t. β :

$$\int \mathcal{N}(y \mid x^t\beta, \gamma^{-1})\, \mathcal{N}(\beta \mid m_n, \gamma^{-1} S_n)\, d\beta$$

↑ Using the general formula at bottom of page 6, we see that this integral is

$$\mathcal{N}\left(y \mid x^t m_n,\ \gamma^{-1}\left(1 + \underset{\substack{\uparrow \\ \text{new point}}}{\mathbf{x}^t} [S_0^{-1} + \underset{\substack{\uparrow \\ \text{matrix of observations}}}{\underline{\underline{X}}^t \underline{\underline{X}}}]^{-1} x\right)\right)$$

x It remains to compute the integral

$$\int \mathcal{N}(y \mid x^t m_n, \gamma^{-1}(1 + x^t[S_0^{-1} + \underline{\underline{X}}^t \underline{\underline{X}}]^{-1} x))\, \text{gamma}(\gamma \mid a_n, b_n)\, d\gamma,$$

which is a Student's t distribution $t(y \mid k, \mu, \tau)$, with

- $k = 2a_n$
- $\mu = x^t m_n$
- $\tau = \frac{a_n}{b_n}\left(1 + x^t [S_0^{-1} + \underline{\underline{X}}^t \underline{\underline{X}}]^{-1} x\right)$

x Summary : → $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \gamma^{-1} I_n)$

→ γ known ⇒ predictive distribution is gaussian, under a gaussian prior on β.

→ γ unknown ⇒ predictive distribution is student, under a gaussian-gamma prior on (β, γ)

Remarks : (i) Conjugate prior on (β, σ^2).

Sometimes it is convenient to work with the variance σ^2 directly, instead of the precision $\gamma = 1/\sigma^2$:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

The conjugate prior on (β, σ^2) is

$$f(\beta, \sigma^2) = \mathcal{N}(\beta \mid m_0, \sigma^2 S_0)\, Ig(\sigma^2 \mid a_0, b_0),$$

where

$Ig(x \mid \alpha, \beta)$ denotes the inverse-gamma distribution, with pdf:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}}, \quad x > 0$$

↖ with

$$EX = \frac{\beta}{\alpha - 1}, \quad \alpha > 1$$

$$\text{Var}\ X = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2$$

$$X \sim \text{gamma}(x \mid \alpha, \beta) \iff \frac{1}{X} \sim Ig(x \mid \alpha, \beta)$$

where we recall that

$$\text{gamma}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$

The posterior distribution of $\beta, \sigma^2 \mid \mathcal{D}_n$ is then

$$f(\beta, \sigma^2 \mid \mathcal{D}_n) = \mathcal{N}(\beta \mid m_n, \sigma^2 S_n)\, Ig(\sigma^2 \mid a_n, b_n),$$

where m_n, S_n, a_n, b_n are given on page 10.

(ii) Noninformative prior on (β, γ) [/ on (β, σ^2)]

Obtained for $(\beta, \gamma) \sim \frac{1}{\gamma}$ $\left[(\beta, \sigma^2) \sim \frac{1}{\sigma^2}\right]$

↳ Special case of the conjugate normal-gamma distribution with $m_0, S_0^{-1} \to 0$, $a_0 = -\frac{d}{2}$, $b_0 \to 0$.

Indeed, take $m_0 = 0$ $\quad a_0 = -\frac{d}{2}$

$S_0^{-1} = \varepsilon I \quad b_0 = \varepsilon$.

Then

$$f(\beta, \gamma) = \mathcal{N}(\beta \mid 0, \gamma^{-1}\varepsilon^{-1} I)\, \text{gamma}(\gamma \mid -\tfrac{d}{2}, \varepsilon)$$

conjugate prior

$$\propto \frac{1}{|\gamma^{-1}\varepsilon^{-1} I|^{1/2}} \underbrace{\exp\left\{-\frac{1}{2}\gamma\varepsilon\beta^t\beta\right\}}_{\downarrow \text{ as } \varepsilon \to 0 \atop 1} \varepsilon^{-\frac{d}{2}} \gamma^{-\frac{d}{2}-1} \underbrace{e^{-\varepsilon\gamma}}_{\downarrow \atop 1}$$

$$\sim \gamma^{\frac{d}{2}} \varepsilon^{\frac{d}{2}} \varepsilon^{-\frac{d}{2}} \gamma^{-\frac{d}{2}-1} \quad \text{as } \varepsilon \to 0$$

$$\sim \gamma^{-1}$$

⇒ The posterior distribution of $(\beta, \gamma \mid \mathcal{L}_n)$ given a noninformative prior $(\beta, \gamma) \sim \frac{1}{\gamma}$ is

$$f(\beta, \gamma \mid \mathcal{L}_n) = \mathcal{N}(\beta \mid (X^tX)^{-1}X^t y, \gamma^{-1}(X^tX)^{-1}) \times \text{gamma}\left(\gamma \,\Big|\, \frac{n-d}{2}, \frac{n-d}{2} s^2\right),$$

with

$$s^2 = \frac{1}{n-d}(y - X\hat{\beta})^t(y - X\hat{\beta}) = \frac{1}{n-d} y^t(I - H)y$$

$\hat{\beta} = (X^tX)^{-1}X^t y$ = LS estimate

$H = X(X^tX)^{-1}X^t$ = projection matrix

since for this choice of m_0, S_0, a_0, b_0, we get

$$S_n = (X^tX)^{-1}$$

$$m_n = (X^tX)^{-1}X^t y = \hat{\beta}$$

$$a_n = \frac{1}{2}(n-d)$$

$$b_n = \frac{1}{2}(y^t y - \hat{\beta}^t(X^tX)\hat{\beta}) = \frac{1}{2} y^t(I - H)y.$$

- The posterior predictive distribution is multivariate t :

$$y \mid \mathcal{L}_n, \underset{\substack{\uparrow \\ \text{one observation } \in \mathbb{R}^d}}{x_0} \sim t(y \mid n-d, \underset{\substack{\uparrow \\ \text{location}}}{x_0^t\hat{\beta}}, s^2(1 + \underset{\substack{\uparrow \\ \text{scale}}}{x_0^t(X^tX)^{-1}x_0}))$$

$$y \mid \mathcal{L}_n, \underset{\substack{\uparrow \\ m\text{-new observations } \in \mathbb{R}^{m\times d}}}{X_0} \sim t(y \mid n-d, X_0\hat{\beta}, s^2(I + X_0(X^tX)^{-1}X_0^t))$$

see page 12, where the scale parameter above corresponds to $1/\tau$.

- Likewise, a non-informative prior on $(\beta, \sigma^2) \sim \frac{1}{\sigma^2}$ yields the posterior

$$(\beta, \sigma^2 \mid \mathcal{L}_n) \sim \mathcal{N}(\beta \mid \hat{\beta}, \sigma^2(X^tX)^{-1})\, \text{Inv-}\chi^2(\sigma^2 \mid n-d, s^2),$$

where

$\text{Inv-}\chi^2(x \mid \nu, s^2)$ denotes the scaled-inverse χ^2 distr; with pdf

$$f(x) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^\nu x^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu s^2}{2x}}, \quad x > 0$$

with $EX = \frac{\nu}{\nu-2} s^2$

$\text{Var}\, X = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4$

Indeed, the posterior distribution $\gamma \mid \mathcal{L}_n$ given on page 15 can be rewritten,

$$\gamma \mid \mathcal{L}_n \sim \text{gamma}\left(\gamma \mid \frac{n-d}{2}, \frac{n-d}{2} s^2\right)$$

$\Leftrightarrow$

$$\sigma^2 = \frac{1}{\gamma} \mid \mathcal{L}_n \sim \mathcal{I}g\left(\sigma^2 \mid \frac{n-d}{2}, \frac{n-d}{2} s^2\right),$$

whose density is precisely Inv-$\chi^2(\sigma^2 \mid n-d, s^2)$, and is given by

$$\frac{\left(\frac{n-d}{2}\right)^{\frac{n-d}{2}}}{\Gamma\left(\frac{n-d}{2}\right)} s^{n-d} (\sigma^2)^{-\left(\frac{n-d}{2}+1\right)} e^{-\frac{(n-d)s^2}{2\sigma^2}}.$$

(iii) Sampling from the posterior distribution.

To draw a sample y from its posterior predictive distribution, either use its analytical expression page 16, or

- first draw $\beta, \gamma \mid \mathcal{L}_n$ from its posterior distribution
- then draw $y \sim \mathcal{N}(\beta X, \gamma^{-1} I)$.

since $f(y \mid \mathcal{L}_n, x) = \iint \underbrace{f(y, \beta, \gamma \mid \mathcal{L}_n, x)} \, d\beta \, d\gamma$

$$= f(y \mid \beta, \gamma, x) \underbrace{f(\beta, \gamma \mid \mathcal{L}_n)}_{\text{posterior}}$$

Obtain B samples $\{\beta^{(b)}, \gamma^{(b)}\}$, $b = 1, .., B$ from the posterior to get a sample $\{y^{(b)}\}$ from the posterior, which amounts to obtaining a sample $\{\beta^{(b)}, \gamma^{(b)}, y^{(b)}\}$ & discarding the params thus marginalizing.

I.4. Evidence Approximation

- Assume that $Y = X\beta + \varepsilon$, $\quad \varepsilon \sim \mathcal{N}(0, \gamma^{-1} I_n)$, $\beta \sim \mathcal{N}(0, \alpha^{-1} I_d)$
- Hyperparameters α and γ are now treated as random, with joint prior $f(\alpha, \gamma)$. This approach is known as Empirical Bayes (EB), Evidence Approximation, or type II Maximum Likelihood.
- Predictive distribution is

$$f(y \mid \mathcal{L}_n, x) = \iiint f(y, \beta, \alpha, \gamma \mid \mathcal{L}_n, x) \, d\beta \, d\alpha \, d\gamma$$

$$= \iiint \underline{f(y \mid \beta, \gamma) f(\beta \mid \mathcal{L}_n, \alpha, \gamma)} \, \underline{f(\alpha, \gamma \mid \mathcal{L}_n)} \, d\beta \, d\alpha \, d\gamma$$

Compare with the expression on page 6: the product of the first two terms is precisely $\mathcal{N}(y \mid x^t \beta, \gamma^{-1}) \mathcal{N}(\beta \mid m_n, S_n)$, where m_n, S_n are given on page 3, while $f(\alpha, \gamma \mid \mathcal{L}_n)$ represents the posterior distribution on the hyperparameters:

$$f(\alpha, \gamma \mid \mathcal{L}_n) \propto \underbrace{f(\mathcal{L}_n \mid \alpha, \gamma)}_{\text{"marginal likelihood"}} \underbrace{f(\alpha, \gamma)}_{\text{prior}}$$

As n gets larger, the posterior is more and more peaked around $(\hat{\alpha}, \hat{\gamma}) = \underset{(\alpha, \gamma)}{\text{argmax}} \, f(\mathcal{L}_n \mid \alpha, \gamma)$. We may substitute $(\hat{\alpha}, \hat{\gamma})$ back into the expression of the predictive distribution, and obtain the approximation

$$f(y \mid \mathcal{L}_n, x) \approx \int f(y \mid \beta, \hat{\gamma}) f(\beta \mid \mathcal{L}_n, \hat{\alpha}, \hat{\gamma}) \, d\beta$$

$$\Rightarrow \boxed{\begin{array}{l} f(y \mid \mathcal{L}_n, x) \approx \mathcal{N}(y \mid m_n^t x,\ \hat{\gamma}^{-1} + x^t S_n x), \\ \text{with} \quad m_n = \hat{\gamma} S_n X^t y, \quad S_n = \hat{\alpha} I_d + \hat{\gamma} X^t X \end{array}}$$

The hyperparameters α and γ are selected from $\mathcal{L}_n$ directly. No need for cross-validation here; hence the name type II ML: $(\hat{\alpha}, \hat{\gamma})$ maximize the marginal likelihood $f(\mathcal{L}_n \mid \alpha, \gamma)$. It remains to find their value.

$$f(\mathcal{L}_n \mid \alpha, \gamma) = \int \underbrace{f(\mathcal{L}_n \mid \beta, \alpha, \gamma)}_{= \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^t \beta, \gamma^{-1})} \underbrace{f(\beta \mid \alpha, \gamma)}_{= \mathcal{N}(\beta \mid 0, \alpha^{-1} I_d)} d\beta$$

The integral can be evaluate using the same formula as given at the bottom of page 6. After calculations, we get

$$f(\mathcal{L}_n \mid \alpha, \gamma) = \mathcal{N}(\mathcal{L}_n \mid 0, \gamma^{-1} I_n + \alpha^{-1} X X^t).$$

$$\log f(\mathcal{L}_n \mid \alpha, \gamma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \boxed{|\gamma^{-1} I_n + \alpha^{-1} X X^t|} - \frac{1}{2} \boxed{y^t (\gamma^{-1} I_n + \alpha^{-1} X X^t)^{-1} y}.$$

$$= \gamma^{-n} |I_n + \frac{\gamma}{\alpha} X X^t|$$

$$\| \quad \text{(C.14) page 697 in Bishop}$$

$$\gamma^{-n} \alpha^{-d} |\alpha I_d + \gamma X^t X|$$

$$\|$$

$$\gamma^{-n} \alpha^{-d} |A|$$

where $A := \alpha I_d + \gamma X^t X$

$$\downarrow \text{(C.7)}$$

$$y^t(\gamma I_n - \gamma X [\alpha I_d + \gamma X^t X]^{-1} X^t \gamma) y$$

$$\|$$

$$\gamma y^t y - \gamma^2 y^t X A^{-1} X^t y$$

$$\|$$

$$\gamma y^t y - \hat{\beta}^t A \hat{\beta} \quad \hat{\beta} = \gamma A^{-1} X^t y$$

Note that $\hat{\beta} = (X^t X + \frac{\alpha}{\gamma} I_d)^{-1} X^t y$; the ridge solution.

$$\Rightarrow \log f(\mathcal{L}_n \mid \alpha, \gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\gamma + \frac{d}{2}\log\alpha - \frac{1}{2}\log|A| \boxed{- \frac{\gamma}{2} y^t y + \frac{1}{2}\hat{\beta}^t A \hat{\beta}}.$$

$$= \frac{1}{2}(\gamma y^t y - \hat{\beta}^t A \beta)$$

$$= \frac{1}{2}(\gamma y^t y - 2\hat{\beta}^t A \hat{\beta} + \hat{\beta}^t A \hat{\beta})$$

$$= \frac{1}{2}(\gamma y^t y - 2\hat{\beta}^t A A^{-1} X^t y \gamma + \hat{\beta}^t (\alpha I + \gamma X^t X) \hat{\beta})$$

$$= \frac{1}{2}(\gamma y^t y - 2\hat{\beta}^t X^t y \gamma + \alpha \hat{\beta}^t \hat{\beta} + \gamma \hat{\beta}^t X^t X \hat{\beta})$$

$$= \frac{1}{2}(\gamma \| y - X\hat{\beta} \|^2 + \alpha \| \hat{\beta} \|^2)$$

$$= \frac{\gamma}{2}(\| y - X\hat{\beta} \|^2 + \frac{\alpha}{\gamma} \| \hat{\beta} \|^2)$$

$= \frac{\gamma}{2} \times$ Penalized Residual Sum of Squares of the Ridge Solution μ_n.

$$\boxed{\begin{array}{l} \log f(\mathcal{L}_n \mid \alpha, \gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\gamma + \frac{d}{2}\log\alpha - \frac{1}{2}\log|A| \\ \qquad - \frac{\gamma}{2} \| y - X\hat{\beta} \|^2 - \frac{\alpha}{2} \| \hat{\beta} \|^2, \\ \text{where } \hat{\beta} = (X^t X + \frac{\alpha}{\gamma} I_d)^{-1} X^t y \end{array}}$$

MARGINAL LIKELIHOOD

We are looking for $(\hat{\alpha}, \hat{\gamma})$ maximizing the marginal likelihood
↳ $\mathcal{L}_n$ is used to estimate hyperparameters → Type II ML.

vs [$\mathcal{L}_n$ being used to estimate parameter β → ML]

↳ First, consider the maximization aith respect to α.

Let (l_i, v_i) = eigenvalue-eigenvector pairs of $\gamma X^t X$.

Since $A = \alpha I_d + \gamma X^t X$, A has eigenvalue-eigenvector pairs $(\alpha + l_i, v_i)$.

Moreaver, $|A| = \prod_{i=1}^{d} (l_i + \alpha) \Rightarrow \log |A| = \sum_{i=1}^{d} \log(\alpha + l_i)$,

and $\frac{d}{d\alpha} \log |A| = \sum_{i=1}^{d} \frac{1}{\alpha + l_i}$.

$$\Rightarrow \frac{\partial}{\partial \alpha} \log f(\mathcal{D}_n | \alpha, \gamma) = \frac{d}{2\hat{\alpha}} - \frac{1}{2} \hat{\beta}^t \hat{\beta} - \frac{1}{2} \sum_{i=1}^{d} \frac{1}{l_i + \hat{\alpha}} = 0$$

$\hat{\beta}$ depends on α.
[we neglected the derivative of $\hat{\beta}$ with respect to α]

$$\boxed{\hat{\alpha} \hat{\beta}^t \hat{\beta} = d - \hat{\alpha} \sum_{i=1}^{d} \frac{1}{l_i + \hat{\alpha}} = \sum_{i=1}^{d} \frac{l_i}{l_i + \hat{\alpha}}} \quad (*)$$

$\hat{\alpha}$ satisfies this equation.
It can be solved iteratively.

↳ Next, consider maximization w.r.t. γ.

Since $X^t X v_i = \frac{l_i}{\gamma} v_i =: \lambda_i v_i$, we see that

$$\frac{d l_i}{d\gamma} = \lambda_i = \frac{l_i}{\gamma} \Rightarrow \frac{d}{d\gamma} \log |A| = \frac{1}{\gamma} \sum_{i=1}^{d} \frac{l_i}{l_i + \alpha}$$

$$\Rightarrow \frac{\partial}{\partial \gamma} \log f(\mathcal{D}_n | \alpha, \gamma) = \frac{n}{2\gamma} - \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{\beta}^t x_i)^2 - \frac{1}{2\gamma} \sum \frac{l_i}{l_i + \alpha}$$

Equating to zero

$$\boxed{\frac{1}{\gamma} = \frac{1}{n - \gamma} \sum_{i=1}^{n} (y_i - \hat{\beta}^t x_i)^2}$$

solve recursively as well.

Remark = Evidence Approximation & EM algorithm.

There are close ties between the recursions derived on page 17, and the EM algorithm. Recall that (see UL: CLUSTERING p. 23)

Goal: maximize the log-likelihood $\ell(\theta) = \log f(x | \theta)$

E-step: compute $Q(\theta, \theta^{(m)}) = \mathbb{E}_{f(z | x, \theta^{(m)})} \{ \log f(x, z | \theta) \}$

M-step : $\theta^{(m+1)} = \underset{\theta}{\arg\max}\, Q(\theta, \theta^{(m)})$.

where θ = parameter of interest
x = observed variable
z = latent variable.

Fact:
$\ell(\theta^{(m)}) \leq \ell(\theta^{(m+1)})$.

↳ In our context, we want to maximize the marginal (log) likelihood

$$f(\mathcal{D}_n | \alpha, \gamma) = \int f(\mathcal{D}_n | \beta, \alpha, \gamma) f(\beta | \alpha, \gamma) \, d\beta$$

our latent variable Z

↳ The complete (log) likelihood is:

$$\log f(\mathcal{D}_n, \beta | \alpha, \gamma) = \underbrace{\boxed{\log f(\mathcal{D}_n | \beta, \alpha, \gamma)}}_{\mathcal{N}(y | X\beta, \gamma^{-1} I_n)} + \underbrace{\boxed{\log f(\beta | \alpha, \gamma)}}_{\mathcal{N}(\beta | 0, \alpha^{-1} I_d)}$$

$$= \boxed{\frac{d}{2} \log\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2} \beta^t \beta} + \boxed{\frac{n}{2} \log \frac{\gamma}{2\pi} - \frac{\gamma}{2} \sum_{i=1}^{n} (y_i - \beta^t x_i)^2}$$

↳ **E-step** =

$$\mathbb{E}\left\{\log f(\mathcal{L}_n, \beta \mid \alpha, \gamma) \mid \mathcal{L}_n, \underset{\uparrow}{\alpha^{(m)}}, \underset{\uparrow}{\gamma^{(m)}}\right\}$$

current parameter values.

$$= \frac{d}{2}\log\left(\frac{\alpha}{2\pi}\right) + \frac{n}{2}\frac{\gamma}{2\pi} - \frac{\alpha}{2}\mathbb{E}\{\beta^t\beta \mid \mathcal{L}_n, \alpha^{(m)}, \gamma^{(m)}\}$$

$$- \frac{\gamma}{2}\sum_{i=1}^{n}\mathbb{E}\left\{(y_i - \beta^t x_i)^2 \mid \begin{matrix}\mathcal{L}_n \\ \alpha^{(m)} \\ \gamma^{(m)}\end{matrix}\right\}$$

where

- $\mathbb{E}\{\beta^t\beta \mid \mathcal{L}_n, \alpha^{(m)}, \gamma^{(m)}\} = m_n^t m_n + S_n,$

  with $\begin{cases} m_n = \gamma^{(m)} S_n X^t y \; (= \hat{\beta}) \\ S_n^{-1} = \alpha^{(m)} I_d + \gamma^{(m)} X^t X \end{cases}$, see page 3.

- $\mathbb{E}\{(y_i - \beta^t x_i)^2 \mid \mathcal{L}_n, \alpha^{(m)}, \gamma^{(m)}\}$

$$= y_i^2 - 2 y_i x_i^t m_n + \underbrace{\mathbb{E}\{Tr(x_i x_i^t \beta\beta^t) \mid \text{—''—}\}}_{\begin{gathered} \| \\ Tr\{x_i x_i^t \, \underbrace{\mathbb{E}(\beta\beta^t \mid \text{—''—})}\} \\ \| \\ Tr(x_i x_i^t S_n) + m_n^t x_i x_i^t m_n \end{gathered}}$$

$$= (y_i - m_n^t x_i)^2 + x_i^t S_n x_i$$

$$\Rightarrow \boxed{\begin{aligned} Q(\theta, \theta^{(m)}) = {} & \frac{d}{2}\log\left(\frac{\alpha}{2\pi}\right) + \frac{n}{2}\log\frac{\gamma}{2\pi} - \frac{\alpha}{2}(m_n^t m_n + S_n) \\ & - \frac{\gamma}{2}\sum_{i=1}^{n}\left((y_i - m_n^t x_i)^2 + x_i^t S_n x_i\right) \end{aligned}}$$

↳ **M-step** = • $\dfrac{\partial}{\partial\alpha} Q(\theta, \theta^{(m)}) = \dfrac{d}{2\alpha^{(m+1)}} - \dfrac{1}{2}(m_n^t m_n + S_n) \underset{=0}{}$

$$\Rightarrow \boxed{\alpha^{(m+1)} = \frac{d}{m_n^t m_n + S_n}}$$

• $\dfrac{\partial}{\partial\gamma} Q(\theta, \theta^{(m)}) = \dfrac{n}{2\gamma^{(m+1)}} - \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\left((y_i - m_n^t x_i)^2 + x_i^t S_n x_i\right)$ $= 0$

$$\Rightarrow \boxed{\gamma^{(m+1)} = \frac{n}{\sum_{i=1}^{n} (y_i - m_n^t x_i)^2 + x_i^t S_n x_i}}$$

Now compare the M-step of the EM algorithm for α with recursion (*) page 17: $\hat{\alpha}\, m_n^t m_n = d - \hat{\alpha} \underbrace{\sum_{i=1}^{d} \frac{1}{l_i + \hat{\alpha}}}_{\begin{gathered}\| \\ Tr(S_n)\end{gathered}}$

Re-arranging terms, this is precisely the recursion above for α. ↗

since l_i = eigenvalue of $\gamma X^t X$ and $S_n^{-1} = \alpha I_d + \gamma X^t X$.

II_ BAYESIAN LOGISTIC REGRESSION

We consider the binary classification problem; $X \in \mathbb{R}^d$, $Y \in \{0, 1\}$, and

$$\log\left\{\frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=0 \mid X=x)}\right\} = \beta^t x \quad ; \quad \beta \in \mathbb{R}^d$$

In a Bayesian framework, we put a Gaussian prior on β, and assume that $\beta \sim f(\beta) = \mathcal{N}(\beta \mid m_0, S_0)$

II.1. Posterior distribution.

- The posterior distribution of β given $\mathcal{L}_n$ is proportional to the product $f(\mathcal{L}_n \mid \beta)\, f(\beta)$, so that

$$\log f(\beta \mid \mathcal{L}_n) = \sum_{i=1}^{n} \left(y_i \log \sigma_i + (1-y_i)\log(1-\sigma_i) \right)$$
$$- \frac{1}{2}(\beta - m_0)^t S_0^{-1} (\beta - m_0) + \text{constant in } \beta,$$

where $\sigma_i = \sigma(\beta^t x_i)$, σ = sigmoid function.

- We are looking for a Gaussian approximation of the posterior. Denote it $q(\beta) = \mathcal{N}(\beta \mid m_n, S_n) \approx f(\beta \mid \mathcal{L}_n)$.

Laplace approximation

↘ m_n = value of β maximizing $\log f(\beta \mid \mathcal{L}_n)$
= MAP estimate

↘ S_n is obtained by considering a Taylor expansion of log of $f(\beta \mid \mathcal{L}_n)$ around its mode m_n:

$$\log f(\beta \mid \mathcal{L}_n) \simeq \log f(m_n \mid \mathcal{L}_n) - \frac{1}{2}(\beta - m_n)^t S_n^{-1} (\beta - m_n),$$

since the derivative / gradient of $\log f(\beta \mid \mathcal{L}_n)$ vanishes at its mode m_n.

where $S_n^{-1} = -\nabla_\beta^2 \{-\log f(\beta \mid \mathcal{L}_n)\}$

$$= S_0^{-1} + X^t W X \;; \qquad W = \begin{pmatrix} \sigma_1(1-\sigma_1) & & 0 \\ & \ddots & \\ 0 & & \sigma_n(1-\sigma_n) \end{pmatrix}$$

see p.14 in SL: LINEAR CLASSIFIERS

evaluated at m_n

II.2. Predictive distribution.

The predictive distribution is given by

$$\mathbb{P}(Y=1 \mid \mathcal{L}_n, x) = \int \boxed{\underset{\sigma(\beta^t x)}{\underset{\shortparallel}{\mathbb{P}(Y=1 \mid x, \beta)}}}\, \boxed{\underset{q(\beta) = \mathcal{N}(\beta \mid m_n, S_n)}{\underset{\simeq}{f(\beta \mid \mathcal{L}_n)}}}\, d\beta$$

new point

① Plug-in approximation:

$$\mathbb{P}(Y=1 \mid \mathcal{L}_n, x) \approx \mathbb{P}(Y=1 \mid x, \beta = m_n) = \sigma(m_n^t x)$$

② MC simulations:

Consider independent samples $\beta_i \sim q(\beta)$. Then

$$\mathbb{P}(Y=1 \mid \mathcal{L}_n, x) \approx \frac{1}{M} \sum_{i=1}^{M} \sigma(\beta_i^t x).$$

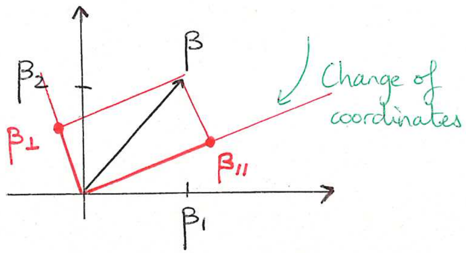
③ Probit approximation:

First, we compute a more tractable expression for the integral

$$\int \sigma(\beta^t x)\, q(\beta)\, d\beta,$$

by expressing $\beta = \begin{pmatrix} \beta_{/\!/} \\ \beta_\perp \end{pmatrix} \begin{matrix} \updownarrow 1 \\ \updownarrow d-1 \end{matrix}$; with $\beta_{/\!/} := \dfrac{\langle \beta, x \rangle}{\sqrt{\langle x, x \rangle}}$

= projection of β onto x



$$\int \sigma(\beta^t x)\, q(\beta)\, d\beta = \iint \sigma(\beta_{/\!/} \|x\|)\, \boxed{q(\beta_\perp \mid \beta_{/\!/})}\, q(\beta_{/\!/})\, d\beta_{/\!/}\, \boxed{d\beta_\perp}$$

integrates to 1

$$\Rightarrow \int \sigma(\beta^t x)\, q(\beta)\, d\beta = \int \sigma(\beta_\parallel \|x\|)\, q(\beta_\parallel)\, d\beta_\parallel$$

Since $\beta \sim$ multivariate Gaussian, $q(\beta_\parallel)$ is $\mathcal{N}(\beta_\parallel \mid \mu_n, \sigma_n^2)$, with

$$\times\ \mu_n := \mathbb{E}(\beta_\parallel \mid \mathcal{D}_n) = \mathbb{E}\left(\frac{x^t\beta}{\|x\|} \,\Big|\, \mathcal{D}_n\right) = \frac{x^t m_n}{\|x\|}, \text{ where}$$

$m_n =$ posterior mean $=$ MAP

$$\times\ \sigma_n^2 := \mathbb{E}\left\{\left(\frac{x^t(\beta - m_n)}{\|x\|}\right)^2 \Big|\, \mathcal{D}_n\right\} \quad \left(= \mathbb{E}\{(\beta_\parallel - \mathbb{E}\beta_\parallel)^2 \mid \mathcal{D}_n\}\right)$$

$$= \frac{x^t}{\|x\|}\, \mathbb{E}\{(\beta - m_n)(\beta - m_n)^t \mid \mathcal{D}_n\}\, x = \frac{x^t S_n x}{\|x\|^2}.$$

Thus, $q(\beta_\parallel) = \mathcal{N}\left(\beta_\parallel \,\Big|\, \frac{x^t m_n}{\|x\|}, \frac{x^t S_n x}{\|x\|^2}\right)$.

$\Rightarrow$ With $s := \beta_\parallel \|x\|$,

$$\int \sigma(\beta_\parallel \|x\|)\, q(\beta_\parallel)\, d\beta_\parallel = \int \sigma(s)\, \underbrace{q\left(\frac{s}{\|x\|}\right)\frac{ds}{\|x\|}}_{\parallel}.$$

density of $\beta_\parallel \|x\| \sim \mathcal{N}(x^t m_n, x^t S_n x)$

We finally get

$$\boxed{\mathbb{P}(Y=1 \mid \mathcal{D}_n, x) = \int \sigma(s)\, \mathcal{N}(s \mid x^t m_n, x^t S_n x)\, ds}$$

The probit approximation of this integral uses Φ in place of σ. Indeed, $\sigma(s) \approx \Phi\left(\sqrt{\frac{\pi}{8}}\, s\right)$, which is easily obtained by equating the slope of the two functions at the origin.

$$\Rightarrow \mathbb{P}(Y=1 \mid \mathcal{D}_n, x) \approx \int \Phi\left(\sqrt{\frac{\pi}{8}}\, s\right) \mathcal{N}(s \mid x^t m_n, x^t S_n x)\, ds,$$

■ which can be analytically computed.

Indeed, consider $X \sim \mathcal{N}(0, \lambda^{-2})$, $Y \sim \mathcal{N}(m, \sigma^2)$, independt.

$$\mathbb{P}(X \leq Y) = \mathbb{E}_Y\, \mathbb{P}(X \leq y) = \int \Phi(\lambda y)\, \mathcal{N}(y \mid m, \sigma^2)\, dy.$$

On the other hand, $X - Y \sim \mathcal{N}(-m, \lambda^{-2} + \sigma^2)$, so that

$$\mathbb{P}(X \leq Y) = \mathbb{P}(X - Y \leq 0) = \Phi\left(\frac{m}{(\lambda^{-2} + \sigma^2)^{1/2}}\right). \quad ■$$

$$\Rightarrow \boxed{\mathbb{P}(Y=1 \mid \mathcal{D}_n, x)} \approx \Phi\left(\frac{x^t m_n}{(\lambda^{-2} + x^t S_n x)^{1/2}}\right); \quad \lambda = \sqrt{\frac{\pi}{8}}$$

use the sigmoid approximation again

$$\approx \boxed{\sigma\left(x^t m_n \left(1 + \frac{\pi}{8}\, x^t S_n x\right)^{-1/2}\right)}$$

Remark: Classify a new observation x as 1 if

$$\mathbb{P}(Y=1 \mid x, \mathcal{D}_n) \geq \mathbb{P}(Y=0 \mid x, \mathcal{D}_n)$$

$$\Updownarrow$$

$$\sigma\left(x^t m_n (\text{---}''\text{---})^{-1/2}\right) \geq 1/2$$

$$\Updownarrow$$

$$x^t m_n \geq 0$$

If the objective is the minimization of the misclassification rate, with equal prior probabilities; then the marginalization over β in the computation of the predictive distribution has no effect.

4 Metropolis-Hastings algorithm.

The previous techniques required the approximation of the posterior distribution. There exists techniques to sample directly from the posterior without requiring to approximate it first. Metropolis-Hastings algorithm (see MS: MCMC) is one of them.

× Goal: to approximate the integral $\int h(\theta) f(\theta)\, d\theta$

× Idea: generate samples $\sim$ density f : $\theta_1, \theta_2, \dots$
the integral is then $\approx \frac{1}{M} \sum_{i=1}^{M} h(\theta_i)$.

× How: start from θ_0, and generate θ_i using a transition kernel, with target density f.

↘ f is known up to a multiplicative constant

↘ choose a proposal density $q(y|\theta)$

and proceed as follows

Given θ_i

(i) generate $y_i \sim q(y|\theta_i)$

(ii) accept/reject

$$\theta_{i+1} = \begin{cases} y_i & \text{w.p. } \rho(\theta_i, y_i) \\ \theta_i & \text{w.p. } 1-\rho(\theta_i, y_i) \end{cases}$$

where

$$\rho(\theta, y) = \min\left\{ \frac{f(y)}{f(\theta)} \frac{q(\theta|y)}{q(y|\theta)}, 1 \right\}$$

METROPOLIS-HASTINGS ALGORITHM

the sequence can take several times the same value (non iid sample)

If $q(\theta|y) = q(y|\theta)$ (symmetric case), always accept points y_i increasing the "likelihood".

Under some general conditions [such as the event $\{\theta_i = \theta_{i+1}\}$ is possible, $q(y|\theta) > 0 \quad \forall (\theta, y)$], then we have

- ergodicity $\frac{1}{M} \sum_{i=1}^{M} h(\theta_i) \longrightarrow \int h(\theta) f(\theta)\, d\theta$
- convergence in total variation. In particular, $\mathbb{P}(\theta_i \in B) \longrightarrow \int_B f(\theta)\, d\theta$

→ In Bayesian logistic Regression, the target density is the posterior

$$f(\beta | \mathcal{L}_n) = \frac{f(\mathcal{L}_n|\beta) f(\beta)}{f(\mathcal{L}_n)}.$$

In MH, we need to compute the ratios,

$$\frac{f(\beta_1|\mathcal{L}_n)}{f(\beta_2|\mathcal{L}_n)} = \frac{f(\mathcal{L}_n|\beta_1) f(\beta_1)}{f(\mathcal{L}_n|\beta_2) f(\beta_2)}, \text{ which is known to us.}$$

In addition, using the symmetric proposal distribution

$$q(\beta_1|\beta_2) = \mathcal{N}(\beta_1 | \beta_2, \xi^2 I_d) = \mathcal{N}(\beta_2 | \beta_1, \xi^2 I_d) = q(\beta_2|\beta_1),$$

we obtain

MH FOR BAYESIAN LOGISTIC REG.

Given β_i

(i) Generate $\gamma_i \sim q(\gamma|\beta_i)$

(ii) Take

$$\beta_{i+1} = \begin{cases} \gamma_i & \text{w.p. } \rho(\beta_i, \gamma_i) \\ \beta_i & \text{w.p. } 1-\rho(\beta_i, \gamma_i) \end{cases}$$

where

$$\rho(\beta, \gamma) = \min\left\{ \frac{f(\mathcal{L}_n|\gamma) f(\gamma)}{f(\mathcal{L}_n|\beta) f(\beta)}, 1 \right\}$$

And compute $\mathbb{P}(Y=1|\mathcal{L}_n, x) \approx \frac{1}{M} \sum_{i=1}^{M} \sigma(\beta_i^t x)$.