VE414 Lecture 14

Jing Liu

UM-SJTU Joint Institute

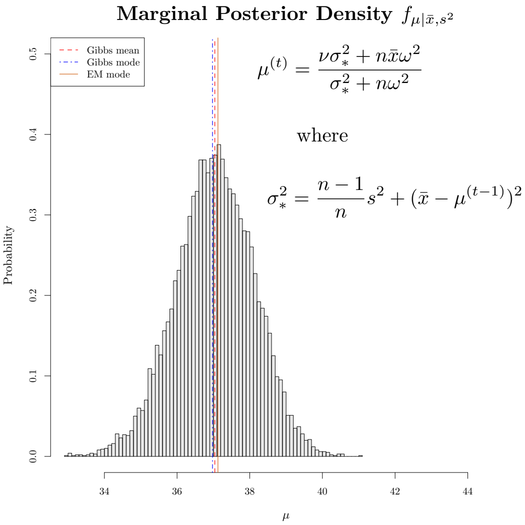
June 27, 2019

- The expectation maximisation (EM) is traditionally used for something else, here it can be used to find the mode of the marginal in a much simpler way.

```
> for (i in 1:9){
+
+    old = (xbar-data_mu[i])^2
+
+    sigma2_star = (n-1)/n * s2 + old
+
+    num = nu * sigma2_star + n * xbar * omega2
+    den = sigma2_star + n * omega2
+
+    data_mu[i+1] = num / den
+
+ }
> data_mu
```

```
 [1]  1.00000 22.64286 31.68842 35.75105 36.89255
 [6] 37.09146 37.12007 37.12404 37.12459 37.12466
```

- In just a few iterations, it produces a value similar to the one from Gibbs.



- In general, given a joint distribution up to a multiplicative constant A,

$$Af_{\mathbf{Y}|X} = Af_{\boldsymbol{\phi},\boldsymbol{\gamma}|X} = q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}$$

where ϕ represents a subset of $\mathbf{Y}$ that we are interested in, i.e.

$$\mathbf{Y} = \begin{bmatrix}\boldsymbol{\phi} & \boldsymbol{\gamma}\end{bmatrix}^{\mathrm{T}}$$

obtaining the marginal is usually impossible, even up to a A^* is also difficult

$$A^* f_{\boldsymbol{\phi}|X} = q_{\boldsymbol{\phi}|X}$$

since it requires either finding the full conditional

$$f_{\boldsymbol{\phi}|X} \propto \frac{q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}}; \qquad f_{\boldsymbol{\phi}|X} \not\propto \frac{q_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}}{q_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}}$$

or evaluating the following integral over the set $\mathcal{D}$ of all possible $\boldsymbol{\gamma}$

$$f_{\boldsymbol{\phi}} \propto \int_{\mathcal{D}} q_{\boldsymbol{\phi},\boldsymbol{\gamma}}(\boldsymbol{\phi},\boldsymbol{\gamma})\, d\boldsymbol{\gamma}$$

- Hence so far using a Monte Carlo method is our only viable option to obtain

$$\hat{\phi}$$

that is, a point estimate of $\phi \mid X$, mean, median or mode.

- The EM algorithm is a way to obtain the mode of the marginal without

$$f_{\boldsymbol{\phi}|X} \qquad \text{or} \qquad q_{\boldsymbol{\phi}|X}$$

in other words, it is an algorithm of maximising the marginal density without knowing the density function or the density function up to a constant!

- Consider the following identity, then logging the both sides, we have

$$f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right) = \frac{f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}\left(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x\right)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x\right)}$$

$$\ln\left(f_{\boldsymbol{\phi}|X}\left(\boldsymbol{\phi} \mid x\right)\right) = \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}\left(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x\right)\right) - \ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x\right)\right)$$

- Taking the expectation on both sides, the term on the left reminds the same

$$\begin{aligned}\mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi}\mid X}\left(\boldsymbol{\phi}\mid x\right)\right)\right] &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi}\mid X}\left(\boldsymbol{\phi}\mid x\right)\right) f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^{*},x\right)\, d\boldsymbol{\gamma} \\ &= \ln\left(f_{\boldsymbol{\phi}\mid X}\left(\boldsymbol{\phi}\mid x\right)\right)\cdot 1\end{aligned}$$

and let the terms on the right become the following

$$\begin{aligned}\alpha\left(\boldsymbol{\phi}\right) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}\left(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x\right)\right)\right] \\ &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}\left(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x\right)\right) f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^{*},x\right)\, d\boldsymbol{\gamma}\end{aligned}$$

$$\begin{aligned}\beta\left(\boldsymbol{\phi}\right) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma}\mid\boldsymbol{\phi},x\right)\right)\right] \\ &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma}\mid\boldsymbol{\phi},x\right)\right) f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^{*},x\right)\, d\boldsymbol{\gamma}\end{aligned}$$

$$\ln\left(f_{\boldsymbol{\phi}\mid X}\left(\boldsymbol{\phi}\mid x\right)\right) = \alpha\left(\boldsymbol{\phi}\right) - \beta\left(\boldsymbol{\phi}\right)$$

the mode $\hat{\boldsymbol{\phi}}$ that maximises $f_{\boldsymbol{\phi}\mid X}$ if and only if $\hat{\boldsymbol{\phi}}$ maximises $\alpha\left(\boldsymbol{\phi}\right) - \beta\left(\boldsymbol{\phi}\right)$.

- Consider the following difference

$$\begin{aligned}
\beta(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi}^*) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)\right)\right] - \mathbb{E}\left[\ln\left(f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)\right)\right] \\
&= \mathbb{E}\left[\ln\left(\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right)\right] \\
&= \int_{\mathcal{D}} \ln\left(\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)\, d\boldsymbol{\gamma}
\end{aligned}$$

- Using Jensen's inequality, we have

$$\begin{aligned}
\beta(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi}^*) &\leq \ln\left(\mathbb{E}\left[\frac{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)}{f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^*, x)}\right]\right) \\
&= \ln\left(\int_{\mathcal{D}} f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)\, d\boldsymbol{\gamma}\right) = 0
\end{aligned}$$

hence increasing/maximising $\alpha(\boldsymbol{\phi})$ increases/maximises $\alpha(\boldsymbol{\phi}) - \beta(\boldsymbol{\phi})$.

Algorithm 1: Expectation-Maximisation

Input : function $f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}$, and $f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}$, initial value $\boldsymbol{\phi}^{(0)}$, tolerance ϵ
Output : mode $\boldsymbol{\phi}_m$

Function EM($f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}$, $f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}$, $\boldsymbol{\phi}^{(0)}$, ϵ):
  $t \leftarrow 1$;
  while $t \leq 1e6$ **do**
    $$\boldsymbol{\phi}^{(t)} \leftarrow \arg\max_{\boldsymbol{\phi}} \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}|X}\left(\boldsymbol{\phi},\boldsymbol{\gamma} \mid x\right)\right) f_{\boldsymbol{\gamma}|\{\boldsymbol{\phi},X\}}\left(\boldsymbol{\gamma} \mid \boldsymbol{\phi}^{(t-1)}, x\right) d\boldsymbol{\gamma}$$
    if $\|\boldsymbol{\phi}^{(t)} - \boldsymbol{\phi}^{(t-1)}\| < \epsilon$ **then**
      $\boldsymbol{\phi}_m \leftarrow \boldsymbol{\phi}^{(t)}$;
      return $\boldsymbol{\phi}_m$; `/* Solution */`
    else
      $t \leftarrow t + 1$;
    end if
  end while
  return *"Warning: 1 million iterations reached without achieving ϵ"* ;
end

- The EM algorithm essentially avoids one of the following two integrals

$$\int_{\mathcal{D}} q_{\boldsymbol{\phi},\boldsymbol{\gamma}}(\boldsymbol{\phi},\boldsymbol{\gamma})\,d\boldsymbol{\gamma} \qquad \text{or} \qquad \int_{\mathcal{D}} q_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi},X)\,d\boldsymbol{\gamma}$$

in return we are required to evaluate with the following integral

$$\begin{aligned}\alpha(\boldsymbol{\phi}) &= \mathbb{E}\left[\ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right)\right] \\ &= \int_{\mathcal{D}} \ln\left(f_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right) f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^*,x)\,d\boldsymbol{\gamma}\end{aligned}$$

Q: Why is this a better deal in general? Because it looks a lot worse!

- Note $\boldsymbol{\phi}^{(t)}$ is the maximiser of α given a specific $\boldsymbol{\phi}^* = \boldsymbol{\phi}^{(t-1)}$ if and only if

$$\boldsymbol{\phi}^{(t)} = \arg\max_{\boldsymbol{\phi}} \int_{\mathcal{D}} \ln\left(q_{\boldsymbol{\phi},\boldsymbol{\gamma}\mid X}(\boldsymbol{\phi},\boldsymbol{\gamma}\mid x)\right) q_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}(\boldsymbol{\gamma}\mid\boldsymbol{\phi}^*,x)\,d\boldsymbol{\gamma}$$

- In addition to the above simplification, when the full conditional distribution $f_{\boldsymbol{\gamma}\mid\{\boldsymbol{\phi},X\}}$ is available, the EM often reduces to simple iterative evaluation.

- In terms of the following model,

$$\begin{aligned} X \mid \{\mu, \sigma^2\} &\sim \mathrm{Normal}\left(\mu, \sigma^2\right) \\ \mu &\sim \mathrm{Normal}\left(\nu, \omega^2\right) \\ \sigma^2 &\sim \varphi_{\sigma^2} \end{aligned}$$

  we have derived the followings last time

$$\begin{aligned} q_{\mu,\sigma^2}\left(\mu, \sigma^2\right) &= \left(\sigma^2\right)^{-(1+n/2)} \cdot \exp\left(-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} - \frac{(\mu-\nu)^2}{2\omega^2}\right) \\ f_{\sigma^2 \mid \{\mu, \bar{x}, s^2\}} &= \mathrm{Scaled\,Inverse}\,\chi^2\left(n, \frac{(n-1)s^2}{n} + (\bar{x}-\mu)^2\right) \end{aligned}$$

- Hence within each iteration, we have to maximise the following w.r.t μ

$$\begin{aligned} \alpha(\mu) &= \mathbb{E}\left[\ln\left(f_{\{\mu,\sigma^2\}\mid\{\bar{x},s^2\}}\left(\mu, \sigma^2 \mid \bar{x}, s^2\right)\right)\right] \\ &= \mathbb{E}\left[-(2+n)\ln\sigma - \frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] - \frac{(\mu-\nu)^2}{2\omega^2} - \ln A \end{aligned}$$

- Rearranging into the following form,

$$\alpha(\mu) = \mathbb{E}\left[-(2+n)\ln\sigma - \frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] - \frac{(\mu-\nu)^2}{2\omega^2} - \ln A$$

$$= -\frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma^2}\right]\left((n-1)s^2 + n(\bar{x}-\mu)^2\right) - \frac{(\mu-\nu)^2}{2\omega^2} \underbrace{-(2+n)\mathbb{E}\left[\ln\sigma\right] - \ln A}_{\text{additive constant w.r.t. } \mu}$$

$$= -\frac{1}{2}\mathbb{E}\left[\frac{1}{\sigma^2}\right]\left((n-1)s^2 + n(\bar{x}-\mu)^2\right) - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant}$$

- Recall the expectation is over σ^2 given $\mu^* = \mu^{(t-1)}$, $\bar{x}$ and s^2, which means

$$\sigma^2 \mid \{\mu^{(t-1)}, \bar{x}, s^2\} \sim \text{Scaled Inverse}\,\chi^2\left(n, \frac{(n-1)s^2}{n} + (\bar{x}-\mu^{(t-1)})^2\right)$$

$$\mathbb{E}\left[\frac{1}{\sigma^2}\right] = \left(\frac{(n-1)s^2}{n} + (\bar{x}-\mu^{(t-1)})^2\right)^{-1}$$

- Thus, in each iteration, we need to solve the following

$$\mu^{(t)} = \arg\max_{\mu} \left\{ \mathbb{E}\left[\ln\left(f_{\{\mu,\sigma^2\}|\{\bar{x},s^2\}}\left(\mu, \sigma^2 \mid \bar{x}, s^2\right)\right)\right]\right\}$$
$$= \arg\max_{\mu} \left\{ -\frac{\left((n-1)s^2 + n(\bar{x}-\mu)^2\right)}{2\sigma_*^2} - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant} \right\}$$

where $\sigma_*^2 = \dfrac{(n-1)s^2}{n} + (\bar{x} - \mu^{(t-1)})^2$.

Q: Have you seen this before?

$$q_\mu \propto \exp\left(-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2} - \frac{(\mu-\nu)^2}{2\omega^2} \right)$$

which is the unnormalised posterior of μ when σ^2 is known and normal prior $\mathrm{Normal}\left(\nu, \omega^2\right)$ is used, the posterior is know to be

$$\mu \mid \{\sigma^2, \bar{x}, s^2\} \sim \mathrm{Normal}\left(\frac{\omega^2\bar{x} + \nu\sigma^2/n}{\omega^2 + \sigma^2/n}, \frac{\omega^2\sigma^2/n}{\omega^2 + \sigma^2/n} \right)$$

- Therefore, the solution to the maximisation in each iteration is simply

$$\mu^{(t)} = \frac{n\omega^2\bar{x} + \nu\sigma_*^2}{n\omega^2 + \sigma_*^2} \qquad \text{where} \quad \sigma_*^2 = \frac{(n-1)s^2}{n} + (\bar{x} - \mu^{(t-1)})^2$$

since the objective function of the maximisation

$$-\frac{\left((n-1)s^2 + n(\bar{x}-\mu)^2\right)}{2\sigma_*^2} - \frac{(\mu-\nu)^2}{2\omega^2} + \text{constant}$$

corresponds to the logarithm of the normal density,

$$\text{Normal}\left(\frac{\omega^2\bar{x} + \nu\sigma_*^2/n}{\omega^2 + \sigma_*^2/n}, \frac{\omega^2\sigma_*^2/n}{\omega^2 + \sigma_*^2/n}\right)$$

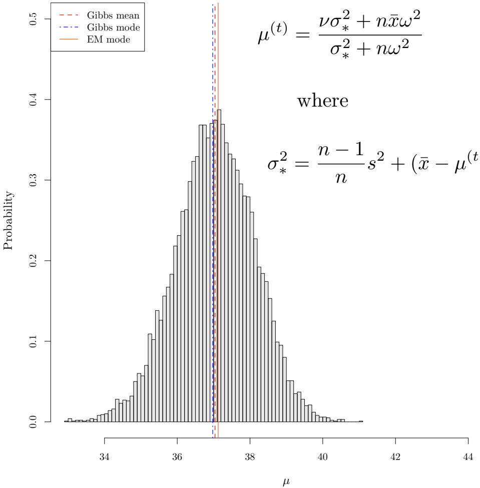
for which we know the maximum happens at where the mean is.

- Using this iterative formula recursively, we reach the the maximiser of

$$f_{\mu|\{\sigma^2,\bar{x},s^2\}}$$

- This leads to what I have used and shown you in the beginning.

Marginal Posterior Density $f_{\mu|\bar{x},s^2}$



$$\mu^{(t)} = \frac{\nu\sigma_*^2 + n\bar{x}\omega^2}{\sigma_*^2 + n\omega^2}$$

where

$$\sigma_*^2 = \frac{n-1}{n}s^2 + (\bar{x} - \mu^{(t-1)})^2$$

- So far we have largely used data to only estimate unobservable,

$$Y$$

- Linear regression model is a way to study the relationship of an observable

$$Y$$

in terms of a set of other observable variables

$$X_1, X_2, \ldots X_k$$

specifically, it is a type of smoothly changing model for

$$f_{Y|\{X_1, X_2, \ldots\}}$$

in which the conditional expectation $\mathbb{E}\left[Y \mid \{X_1, \ldots X_k\}\right]$ has a form that is linear in a set of unobservable β_i, which are often known as the parameters

$$\mathbb{E}\left[Y \mid \{X_1, \ldots X_k\}\right] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}$$

- In addition to being linear,

$$\mathbb{E}\left[Y \mid \{X_1, \ldots, X_k\}\right] = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$$

- The variability around the mean, i.e. the error,

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$$

  is often assumed to be normal

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Normal}\left(0, \sigma^2\right)$$

- Under the above specification, we have the following density function

$$\begin{aligned} f_{\{Y_1, Y_2, \cdots, Y_n\} \mid \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2\}} &= \prod_{i=1}^{n} f_{Y_i \mid \{\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2\}} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right) \end{aligned}$$

- We can put the density function into a vector form,

$$
\begin{aligned}
f_{\{Y_1,Y_2,\cdots,Y_n\}|\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n,\boldsymbol{\beta},\sigma^2\}} &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\mathsf{RSS}\right)
\end{aligned}
$$

where residual sum of squares is given by

$$
\mathsf{RSS} = \sum_{i=1}^{n}\left(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 = \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)
$$

- Thus our model in vector form is $\mathbf{Y} \mid \{\mathbf{X},\boldsymbol{\beta},\sigma^2\} \sim \mathrm{Normal}\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}\right)$.

Q: What would frequentists do next?

- Frequentists would maximise the likelihood by treating the density function as a function of the unknown parameters, which is equivalent to minimise

$$
\mathsf{RSS}\left(\mathbf{b}\right) = \left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)
$$

- Recall to minimise a function,

$$\mathsf{RSS}\left(\mathbf{b}\right) = \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

  we set the gradient to zero,

$$\nabla\mathsf{RSS} = 0 - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

  Setting this to zero, we have

$$\hat{\boldsymbol{\beta}}_{\mathsf{MLE}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Hence, the fitted value is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{P}\mathbf{y}$$

  and the residual can be found using

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I} - \mathbf{P}\right)\mathbf{y}$$

- With more linear algebra, we have

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{e}$$

which means it is unbiased as expected,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right] = \boldsymbol{\beta}$$

- The variance is given by

$$\begin{aligned}\mathrm{Var}\left[\hat{\boldsymbol{\beta}} \mid \mathbf{X}\right] &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{Var}\left[\boldsymbol{\varepsilon} \mid \mathbf{X}\right]\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned}$$

- With the normal assumption, we see

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$

- To estimate σ^2, frequentists typically use the following

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}} \qquad \text{where} \quad \hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y}$$

  which is unbiased as well as being consistent.

- It can be shown the residual

$$\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\,\mathbf{y} = (\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})$$

  is an unbiased and consistent estimator of the error $\mathbf{e}$, and the variance is

$$\begin{aligned}\operatorname{Var}[\hat{\mathbf{e}} \mid \mathbf{X}] &= \operatorname{Var}[(\mathbf{I}-\mathbf{P})\,(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \mid \mathbf{X}] \\ &= (\mathbf{I}-\mathbf{P})\operatorname{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} \\ &= (\mathbf{I}-\mathbf{P})\,\sigma^2\mathbf{I}\,(\mathbf{I}-\mathbf{P})^{\mathrm{T}} = \sigma^2\,(\mathbf{I}-\mathbf{P})\end{aligned}$$

- Thus with the normal assumption, we have

$$\hat{\mathbf{e}} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\,(\mathbf{I}-\mathbf{P})\right)$$

Q: How would Bayesian approach the same problem?

$$f_{\mathbf{Y}|\{\mathbf{X},\boldsymbol{\beta},\sigma^2\}} = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\mathsf{RSS}\left(\boldsymbol{\beta}\right)\right)$$

where

$$\mathsf{RSS}\left(\boldsymbol{\beta}\right) = \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

- Consider using a normal prior for $\boldsymbol{\beta} \sim \mathrm{Normal}\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\right)$, then

$$\begin{aligned} f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) &= \frac{1}{\sqrt{(2\pi)^k \det\left(\boldsymbol{\Sigma}_0\right)}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right) \end{aligned}$$

Q: What is the conditional posterior of $\boldsymbol{\beta}$?

$$f_{\boldsymbol{\beta}|\{\sigma^2,\mathbf{Y},\mathbf{X}\}}$$

- Consider using the precision parameter in the likelihood instead of σ^2, that is

$$\tau = \frac{1}{\sigma^2}$$

and using a gamma prior for $\tau \sim \mathrm{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$,

$$\begin{aligned} f_\tau &= \frac{\left(\nu_0\sigma_0^2/2\right)^{\nu_0/2}}{\Gamma\left(\nu_0/2\right)} \tau^{\nu_0/2-1} \exp\left(-\frac{\nu_0\sigma_0^2}{2}\tau\right) \\ &\propto \tau^{\nu_0/2-1} \exp\left(-\frac{\nu_0\sigma_0^2}{2}\tau\right) \end{aligned}$$

Q: What is the conditional posterior of $\boldsymbol{\tau}$?

$$f_{\sigma^2|\{\boldsymbol{\beta},\mathbf{Y},\mathbf{X}\}}$$

Q: How can we sample from the Joint posterior?

$$f_{\{\boldsymbol{\beta},\sigma^2\}|\{\mathbf{Y},\mathbf{X}\}}$$

- Since both conditionals are readily available, and both are pretty standard,

$$\boldsymbol{\beta} \mid \{\sigma^2, \mathbf{Y}, \mathbf{X}\} \sim \mathrm{Normal}\left(\mathbf{m}, \mathbf{V}\right)$$
$$\sigma^2 \mid \{\boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}\} \sim \text{Inverse-Gamma}\left(\alpha, \beta\right)$$

where

$$\begin{aligned}
\mathbf{m} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2\right)^{-1} \left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{y}/\sigma^2\right)^{-1} \\
\mathbf{V} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2\right)^{-1} \\
\alpha &= \frac{\nu_0 + n}{2}; \qquad \beta = \frac{\nu_0\sigma_0^2 + \mathsf{RSS}\left(\boldsymbol{\beta}\right)}{2}
\end{aligned}$$

and positivity is satisfied, using Gibbs sampling is then straightforward

$$\left(\boldsymbol{\beta}, \sigma^2\right) \in \mathbb{R}^k \times (0, \infty)$$

- If other priors are used, we will have a different joint and a different sampling scheme, but the essences of Bayesian linear regression are the same.