Sequential Models in Data Science: Bayesian Regression

Ariel Fuxman & Gilad Zusman

2. Computing the Posterior of w

1. Denote $Y = (y_1, \ldots, y_n)$. Since $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, then $y_i|w \sim \mathcal{N}(y(x_i, w), \sigma^2)$ as a sum of a gaussian variable and a constant variable. Hence, the likelihood of w is:

$$
\begin{aligned}
f(Y|w) = \prod_{i=1}^{n} f(y_i|w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - y(x_i, w)^2\right)\right) = \\
= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y(x_i, w)^2\right)\right).
\end{aligned}
$$

2. The Bayes formula for the posterior of w is:

$$
f(w|Y) = \frac{f(Y|w)\pi(w)}{f(Y)},
$$

where $\pi(w)$ is the joint distribution of the weights w_i.

3. Let us define the matrix $\Phi := \left(\phi_j(x_i)\right)_{i,j=1,\ldots,n}$. Since $y(x_i, w) = \sum_{j=1}^{n} w_j \phi_j(x_i)$, it allows us to use matrix multiplication to write:

$$
\sum_{i=1}^{n}\left(y_i - y(x_i, w)^2\right) = \|Y - \Phi w\|^2 .
$$

Assuming the weights w_i are independent,

$$
\pi(\omega) = \prod_{i=1}^{n} \pi(w_i) = \prod_{i=1}^{n} \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2} w_i^2\right) = \left(\frac{\alpha}{2\pi}\right)^{n/2} \exp\left(-\frac{\alpha}{2}\|w\|^2\right).
$$

The MAP estimator is defined as $w_{\text{MAP}} := \underset{w}{\operatorname{argmax}} f(w|Y)$. Looking at the Bayes formula for the posterior distribution, we see that $w_{\text{MAP}} = \underset{w}{\operatorname{argmax}} f(Y|w)\pi(w)$, since the denominator does not depend on w.

$$f(Y|w)\pi(w) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y(x_i,w)^2\right)\right)\left(\frac{\alpha}{2\pi}\right)^{n/2} \exp\left(-\frac{\alpha}{2}\left\|w\right\|^2\right)$$

$$\Longrightarrow \ln f(Y|w)\pi(w) = \ln\left(\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\left(\frac{\alpha}{2\pi}\right)^{n/2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y(x_i,w)^2\right) - \frac{\alpha}{2}\left\|w\right\|^2.$$

The term $\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\left(\frac{\alpha}{2\pi}\right)^{n/2}$ does not depend on w, so:

$$\begin{aligned}
w_{\text{MAP}} &= \underset{w}{\operatorname{argmax}}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y(x_i,w)^2\right) - \frac{\alpha}{2}\left\|w\right\|^2\right) = \\
&= \underset{w}{\operatorname{argmin}}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y(x_i,w)^2\right) + \frac{\alpha}{2}\left\|w\right\|^2\right) = \\
&= \underset{w}{\operatorname{argmin}}\left(\sum_{i=1}^{n}\left(y_i - y(x_i,w)^2\right) + \alpha\sigma^2\left\|w\right\|^2\right) \\
&= \underset{w}{\operatorname{argmin}}\left(\left\|Y - \Phi w\right\|^2 + \alpha\sigma^2\left\|w\right\|^2\right),
\end{aligned}$$

which is the solution a regularized least squares problem with regularization parameter $\lambda = \alpha\sigma^2$.

3. The Gaussian Nature of the Posterior and the Prediction

1. We will prove the proposition in two ways, the second way relying on the Lemmas. We will show that $w|Y \sim \mathcal{N}(\mu, \Sigma)$, where

$$\begin{aligned}
\mu &= \left(\Phi^T\Phi + \alpha\sigma^2 I_n\right)^{-1}\Phi^T Y \in \mathbb{R}^n, \\
\Sigma &= \sigma^2\left(\Phi^T\Phi + \alpha\sigma^2 I_n\right)^{-1} \in M_{n\times n}\left(\mathbb{R}\right).
\end{aligned}$$

Proof. We saw that

$$f(Y|w) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|Y - \Phi w\right\|^2\right)$$

and

$$\pi(w) \propto \exp\left(-\frac{\alpha}{2}\left\|w\right\|^2\right),$$

so the posterior

$$f(w|Y) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\|Y-\Phi w\|^2+\alpha\|w\|^2\right)\right).$$

We are now left with showing that:

$$\frac{1}{\sigma^2}\|Y-\Phi w\|^2+\alpha\|w\|^2=(w-\mu)^T\Sigma^{-1}(w-\mu)+Const.$$

since that constant (with respect to w) is inside the exponent, and just gets absorbed as a multiplicative constant, still achieving proportionality to the Gaussian $\exp\left(-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\right)$. Indeed, the right side is:

$$\begin{aligned}(w-\mu)^T\Sigma^{-1}(w-\mu) &= \left(w^T-\mu^T\right)\Sigma^{-1}(w-\mu)= \\ &= w^T\Sigma^{-1}w-\underbrace{w^T\Sigma^{-1}\mu}_{\in\mathbb{R}}-\underbrace{\mu^T\Sigma^{-1}w}_{\in\mathbb{R}}+\mu^T\Sigma^{-1}\mu= \\ &= w^T\Sigma^{-1}w-2\mu^T\Sigma^{-1}w+\mu^T\Sigma^{-1}\mu,\end{aligned}$$

and the left side is:

$$\begin{aligned}\frac{1}{\sigma^2}\|Y-\Phi w\|^2+\alpha\|w\|^2 &= \frac{1}{\sigma^2}(Y-\Phi w)^T(Y-\Phi w)+\alpha w^Tw= \\ &= \frac{1}{\sigma^2}\left(Y^T-w^T\Phi^T\right)(Y-\Phi w)+\alpha w^Tw= \\ &= \frac{1}{\sigma^2}\left(Y^TY-\underbrace{Y^T\Phi w}_{\in\mathbb{R}}-\underbrace{w^T\Phi^TY}_{\in\mathbb{R}}+w^T\Phi^T\Phi w\right)+\alpha w^Tw= \\ &= \frac{1}{\sigma^2}\left(Y^TY-2Y^T\Phi w+w^T\Phi^T\Phi w\right)+\alpha w^Tw= \\ &= \frac{1}{\sigma^2}w^T\left(\Phi^T\Phi+\alpha\sigma^2I_n\right)w-\frac{2}{\sigma^2}Y^T\Phi w+\frac{1}{\sigma^2}Y^TY.\end{aligned}$$

To continue, note that:

$$\Sigma=\sigma^2\left(\Phi^T\Phi+\alpha\sigma^2I_n\right)^{-1}\Longrightarrow\Sigma^{-1}=\frac{1}{\sigma^2}\left(\Phi^T\Phi+\alpha\sigma^2I_n\right)$$

In addition, since Σ is symmetric (easy to see),

$$\mu=\left(\Phi^T\Phi+\alpha\sigma^2I_n\right)^{-1}\Phi^TY\Longrightarrow\mu=\frac{1}{\sigma^2}\Sigma\Phi^TY\Longrightarrow\mu^T=\frac{1}{\sigma^2}Y^T\Phi\Sigma^T=\frac{1}{\sigma^2}Y^T\Phi\Sigma$$

which means that:

$$\mu^T \Sigma^{-1} = \frac{1}{\sigma^2} Y^T \Phi.$$

So we have shown that

$$w^T \Sigma^{-1} w - 2\mu^T \Sigma^{-1} w = \frac{1}{\sigma^2} w^T \left(\Phi^T \Phi + \alpha \sigma^2 I_n \right) w - \frac{2}{\sigma^2} Y^T \Phi w.$$

The second way to prove the theorem is using **Lemma 1** and **Lemma 2**: ☐

Looking at the expressions for the probability distributions of w and $Y|w$,

$$w \sim \mathcal{N}(0, \frac{1}{\alpha} I_n)$$

$$Y|w \sim \mathcal{N}(\Phi w, \sigma^2 I_n),$$

so by **Lemma 1** we conclude that:

$$\begin{pmatrix} w \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{\alpha} I & \frac{1}{\alpha} \Phi^T \\ \frac{1}{\alpha} \Phi & \frac{1}{\alpha} \Phi \Phi^T + \sigma^2 I \end{pmatrix} \right).$$

Now, using **Lemma 2**,

$$w|Y \sim \mathcal{N}\left(\frac{1}{\alpha} \Phi^T \left(\frac{1}{\alpha} \Phi \Phi^T + \sigma^2 I \right)^{-1} Y, \frac{1}{\alpha} I - \frac{1}{\alpha} \Phi^T \left(\frac{1}{\alpha} \Phi \Phi^T + \sigma^2 I \right)^{-1} \frac{1}{\alpha} \Phi \right).$$

To show that

$$\mu = \frac{1}{\alpha} \Phi^T \left(\frac{1}{\alpha} \Phi \Phi^T + \sigma^2 I \right)^{-1},$$

$$\Sigma = \frac{1}{\alpha} I - \frac{1}{\alpha} \Phi^T \left(\frac{1}{\alpha} \Phi \Phi^T + \sigma^2 I \right)^{-1} \frac{1}{\alpha} \Phi,$$

we use the **Woodbury formula** (see Wikipedia)**:**

Proposition 1. *Let $A_{n \times n}, C_{k \times k}, U_{n \times k}, V_{k \times n}$ be matrices. Then,*

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + V A^{-1} U \right)^{-1} V A^{-1}$$

Continuing with the proof,

$$
\begin{aligned}
\frac{1}{\alpha}\Phi^T\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1} &= \Phi^T\frac{1}{\alpha}\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1} = \\
&= \Phi^T(\alpha I)^{-1}\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1} = \\
&= \Phi^T\left(\Phi\Phi^T+\alpha\sigma^2 I\right)^{-1} = \\
&= \left(\Phi^{-T}\right)^{-1}\left(\Phi\Phi^T+\alpha\sigma^2 I\right)^{-1} = \\
&= \left(\Phi+\alpha\sigma^2\Phi^{-T}\right)^{-1} = \\
&= \left(\Phi^{-T}\Phi^T\Phi+\alpha\sigma^2\Phi^{-T}\right)^{-1} = \\
&= \left(\Phi^{-T}\left(\Phi^T\Phi+\alpha\sigma^2 I\right)\right)^{-1} = \\
&= \left(\Phi^T\Phi+\alpha\sigma^2 I\right)^{-1}\Phi^T,
\end{aligned}
$$

showing that

$$\mu = \frac{1}{\alpha}\Phi^T\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1} Y.$$

Now, using **Proposition 1**,

$$
\begin{aligned}
\frac{1}{\alpha}I-\frac{1}{\alpha}\Phi^T\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1}\frac{1}{\alpha}\Phi &= \frac{1}{\alpha}\left(I-\Phi^T\frac{1}{\alpha}\left(\frac{1}{\alpha}\Phi\Phi^T+\sigma^2 I\right)^{-1}\Phi\right) = \\
&= \frac{1}{\alpha}\left(I-\left(\Phi^T\Phi+\alpha\sigma^2 I\right)^{-1}\Phi^T\Phi\right) = \\
&= \frac{1}{\alpha}\left(I-\left(\Phi^T\Phi+\alpha\sigma^2 I\right)^{-1}\left(\left(\Phi^T\Phi\right)^{-1}\right)^{-1}\right) = \\
&= \frac{1}{\alpha}\left(I-\left(I+\alpha\sigma^2\left(\Phi^T\Phi\right)^{-1}\right)^{-1}\right) = \\
&= \frac{1}{\alpha}\left(I-I^{-1}+I^{-1}\alpha\sigma^2\left(\Phi^T\Phi\right)^{-1}\left(I+\alpha\sigma^2\left(\Phi^T\Phi\right)^{-1}\right)^{-1}\right) = \\
&= \frac{1}{\alpha}\left(\alpha\sigma^2\left(\Phi^T\Phi+\alpha\sigma^2 I\right)^{-1}\right) = \Sigma
\end{aligned}
$$

2. In the previous question, we saw that $w|Y\sim\mathcal{N}(\mu,\Sigma)$. Note that:

$$y(x_\star,w)=\sum_{i=1}^{n} w_i\phi_i(x_\star)=\left\langle\left(\phi_1(x_\star),\ldots,\phi_n(x_\star)\right),w\right\rangle = f^T w.$$

This means that $y_\star|w,Y=y_\star|w\sim\mathcal{N}(f^T w,\widehat{\sigma}^2)$ (Since the weights alone entirely determine the distrubution of w). Using **Lemma 1**, we conclude that:

$$y_\star | Y \sim \mathcal{N}(\underbrace{f^T \mu}_{y(x_\star, \mu)}, f^T \Sigma f + \widehat{\sigma}^2).$$

4. Sparse Bayseian Learning

1. As expected, the proof is very similar to **4.1**. This time,

$$\begin{aligned}
\pi(\omega) &= \prod_{i=1}^{n} \pi(w_i) = \prod_{i=1}^{n} \sqrt{\frac{\alpha_i}{2\pi}} \exp\left(-\frac{\alpha_i}{2} w_i^2\right) = (2\pi)^{n/2} \left(\prod_{i=1}^{n} \alpha_i\right)^{1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \alpha_i w_i^2\right) = \\
&= (2\pi)^{n/2} \left(\det(A)\right)^{1/2} \exp\left(-\frac{1}{2} w^T A w\right) \propto \exp\left(-\frac{1}{2} w^T A w\right).
\end{aligned}$$

Meaning that this time have to show that:

$$\beta \left\| Y - \Phi w \right\|^2 + w^T A w = (w - m)^T \Sigma^{-1} (w - m) + Const,$$

where:

$$\begin{aligned}
m &= \beta \Sigma \Phi^T Y \in \mathbb{R}^n, \\
\Sigma &= \left(\beta \Phi^T \Phi + A\right)^{-1} \in M_{n \times n}(\mathbb{R}).
\end{aligned}$$

This time, the left hand side is:

$$w^T \underbrace{\left(\beta \Phi^T \Phi + A\right)}_{\Sigma^{-1}} w - 2 \underbrace{\beta Y^T \Phi}_{m} w + \beta Y^T Y$$

and the right side is unchanged, finishing the proof.

2.

$$Y | w, \alpha, \beta \sim \mathcal{N}\left(\Phi w, \sigma^2 I\right),$$

$$w | \alpha, \beta \sim \mathcal{N}\left(0, A^{-1}\right),$$

so by **Lemma 1**,

$$Y | \alpha, \beta \sim \mathcal{N}\left(0, \Phi A^{-1} \Phi^T + \sigma^2 I\right).$$

To continue, we use **Weinstein–Aronszajn identity** (see Wikipedia):

Proposition 2. *Let $A_{m\times n}, B_{n\times m}$ be matrices. Then,*

$$\det\left(I_m + AB\right) = \det\left(I_n + BA\right)$$

We use it to derive the expression for the determinant of the covariance matrix:

$$\begin{aligned}|A|\left|\Phi A^{-1}\Phi^T + \sigma^2 I\right| &= \left|\sigma^2 I\right| |A| \left|(\beta\Phi)\left(A^{-1}\Phi^T\right) + I\right| = \\ &= \left|\sigma^2 I\right| |A| \left|A^{-1}\beta\Phi^T\Phi + I\right| \\ &= \left(\frac{1}{\beta}\right)^n \left|\underbrace{\beta\Phi\Phi^T + A}_{\Sigma^{-1}}\right|\end{aligned}$$

$$\Longrightarrow \left|\Phi A^{-1}\Phi^T + \sigma^2 I\right| = \left(\frac{1}{\beta}\right)^n \frac{1}{|A|}\left|\Sigma^{-1}\right|$$

$$\Longrightarrow \ln\left|\Phi A^{-1}\Phi^T + \sigma^2 I\right| = -\left|\Sigma\right| - \sum_{i=1}^{n}\ln\left(\alpha_i\right) - n\ln\beta.$$

Using **Proposition 1**,

$$\left(\Phi A^{-1}\Phi^T + \sigma^2 I\right)^{-1} = \beta I - \beta\Phi\underbrace{\left(A + \beta\Phi^T\Phi\right)^{-1}}_{\Sigma}\Phi^T\beta$$

$$\Longrightarrow Y^T\left(\Phi A^{-1}\Phi^T + \sigma^2 I\right)^{-1} Y = \beta Y^T Y - \underbrace{\beta Y^T\Phi\Sigma\Phi^T\beta Y}_{m^T\Sigma^{-1}m}$$

since,

$$m^T\Sigma^{-1}m = \left(\beta\Sigma\Phi^T Y\right)^T\Sigma^{-1}\beta\Sigma\Phi^T Y = \beta Y^T\Phi\underbrace{\Sigma^T}_{\Sigma}\Sigma^{-1}\beta\Sigma\Phi^T Y = \beta Y^T\Phi\Sigma\Phi^T\beta Y.$$

Lastly,

$$f(Y|\alpha,\beta) = (2\pi)^{-n/2}\left(\left|\Phi A^{-1}\Phi^T + \sigma^2 I\right|\right)^{-1/2}\exp\left(-\frac{1}{2}Y^T\left(\Phi A^{-1}\Phi^T + \sigma^2 I\right)^{-1} Y\right)$$

$$\ln f(Y|\alpha,\beta) = \frac{1}{2}\left(-n\ln\left(2\pi\right) + \ln|\Sigma| + \sum_{i=1}^{n}\ln\alpha_i + n\ln\beta - \beta Y^T Y + m^T\Sigma^{-1}m\right).$$