1 MAP with Colored Noise

Recall the ordinary least squares (OLS) model. We have a dataset $\mathcal{D} = \{(\vec{a}_i, y_i)\}_{i=1}^n$ and assume that each $y_i$ is a linear function of $\vec{a}_i$, plus some independent Gaussian noise, which we have rescaled to have variance 1:

$$z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$$

$$y_i = \vec{a}_i^\top \vec{w} + z_i$$

Initially we used the geometric interpretation of OLS to solve for $\vec{w}$. The previous two lectures showed how we can find $\vec{w}$ with estimators instead:

1. Maximum likelihood estimation (MLE):

$$\hat{\vec{w}} = \arg \max_{\vec{w}} \log P(\mathcal{D} | \vec{w})$$

2. Maximum a posteriori estimation (MAP):

$$\hat{\vec{w}} = \arg \max_{\vec{w}} \log P(\vec{w} | \mathcal{D}) = \arg \max_{\vec{w}} \log P(\mathcal{D} | \vec{w}) + \log P(\vec{w})$$

When deriving ridge regression via MAP estimation, our prior assumed that $w_i$ were i.i.d. (univariate) Gaussian, but more generally, we can allow $\vec{w}$ to be any multivariate Gaussian:

$$\vec{w} \sim \mathcal{N}(\vec{\mu}_w, \Sigma_w)$$

Recall (see Discussion 4) that we can rewrite a multivariate Gaussian variable as an affine transformation of a standard Gaussian variable:

$$\vec{w} = \Sigma_w^{1/2} \vec{v} + \vec{\mu}_w$$

$$\vec{v} \sim \mathcal{N}(0, I)$$

This change of variable is sometimes called the reparameterization trick.

Plugging this reparameterization into our approximation $\vec{Y} \approx A\vec{w}$ gives

$$\vec{Y} \approx A\Sigma_w^{1/2} \vec{v} + A\vec{\mu}_w$$

$$A\Sigma_w^{1/2} \vec{v} \approx \vec{Y} - A\vec{\mu}_w$$

$$\hat{\vec{v}} = (\Sigma_w^{1/2} A^\top (A\Sigma_w^{1/2} + I)^{-1} \Sigma_w^{1/2} A)^{-1} (\vec{Y} - A\vec{\mu}_w)$$
Since the variance from data and prior have both been normalized, the noise-to-signal ratio ($\lambda$) is equal to 1.

However $\vec{v}$ is not what we care about – we need to convert back to the actual weights $\vec{w}$ in order to make predictions. Using our identity again,

$$\hat{\vec{w}} = \bar{\mu}_w + \Sigma_w^{1/2}(\Sigma_w^{1/2}A^\top A\Sigma_w^{1/2} + I)^{-1}\Sigma_w^{1/2}A^\top (\vec{y} - A\bar{\mu}_w)$$

$$= \bar{\mu}_w + (A^\top A + \sum_w^{-1/2}\Sigma_w^{-1/2})^{-1}A^\top (\vec{y} - A\bar{\mu}_w)$$

Note that there are two terms: the prior mean $\bar{\mu}_w$, plus another term that depends on both the data and the prior. The precision matrix of $\vec{w}$'s prior ($\Sigma_w^{-1}$) controls how the data fit error affects our estimate.

To gain intuition, let us consider the simplified case where

$$\Sigma_w = \begin{bmatrix} \sigma_{w,1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{w,2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{w,n}^2 \end{bmatrix}$$

When the prior variance $\sigma_{w,j}^2$ for dimension $j$ is large, the prior is telling us that $w_j$ may take on a wide range of values. Thus we do not want to penalize that dimension as much, preferring to let the data fit sort it out. And indeed the corresponding entry in $\Sigma_w^{-1}$ will be small, as desired.

Conversely if $\sigma_{w,j}^2$ is small, there is little variance in the value of $w_j$, so $w_j \approx \mu_j$. As such we penalize the magnitude of the data-fit contribution to $\hat{w}_j$ more heavily.

1.1 Alternative derivation

MAP with colored noise can be expressed as:

$$\vec{u}, \vec{v} \text{ i.i.d. } \mathcal{N}(0, I) \tag{3}$$

$$\begin{bmatrix} \vec{Y} \\ \vec{w} \end{bmatrix} = \begin{bmatrix} R_z & AR_w \\ 0 & R_w \end{bmatrix} \begin{bmatrix} \vec{u} \\ \vec{v} \end{bmatrix} \tag{4}$$

where $R_z$ and $R_w$ are relationships with $w$ and $z$, respectively. Note that the $R_w$ appears twice because our model assumes $\vec{Y} = A\vec{w} +$ noise, so if $\vec{w} = R_w\vec{v}$, then we must have $\vec{Y} = AR_w\vec{v} +$ noise.

We want to find the posterior $\vec{w} \mid \vec{Y}$. The formulation above makes it relatively easy to find the posterior of $\vec{Y}$ conditioned on $\vec{w}$ (see below), but not vice-versa. So let’s pretend instead that

$$\vec{u}', \vec{v}' \text{ i.i.d. } \mathcal{N}(0, I)$$

$$\begin{bmatrix} \vec{w} \\ \vec{Y} \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D \end{bmatrix} \begin{bmatrix} \vec{u}' \\ \vec{v}' \end{bmatrix}$$
Now $\tilde{w} \mid \tilde{Y}$ is straightforward. Since $v' = D^{-1} \tilde{Y}$, the conditional mean and variance of $\tilde{w} \mid \tilde{Y}$ can be computed as follows:

$$
\mathbb{E}[\tilde{w} \mid \tilde{Y}] = \mathbb{E}[Au' + Bv' \mid \tilde{Y}]
= \mathbb{E}[Au' \mid \tilde{Y}] + \mathbb{E}[BD^{-1}\tilde{Y} \mid \tilde{Y}]
= A \mathbb{E}[u'] + \mathbb{E}[BD^{-1}\tilde{Y} \mid \tilde{Y}]
= BD^{-1}\tilde{Y}
$$

$$
\text{var}(\tilde{w} \mid \tilde{Y}) = \mathbb{E}[(\tilde{w} - \mathbb{E}[\tilde{w}])(\tilde{w} - \mathbb{E}[\tilde{w}])^\top \mid \tilde{Y}]
= \mathbb{E}[(Au' + BD^{-1}\tilde{Y} - BD^{-1}\tilde{Y})(Au' + BD^{-1}\tilde{Y} - BD^{-1}\tilde{Y})^\top \mid \tilde{Y}]
= \mathbb{E}[(Au')(Au')^\top \mid \tilde{Y}]
= \mathbb{E}[Au'(u')^\top A^\top]
= A \mathbb{E}[u'(u')^\top A^\top]
= \text{var}(u') = I
= AA^\top
$$

In both cases above where we drop the conditioning on $\tilde{Y}$, we are using the fact $u'$ is independent of $v'$ (and thus independent of $\tilde{Y} = Dv'$). Therefore

$$
\tilde{w} \mid \tilde{Y} \sim \mathcal{N}(BD^{-1}\tilde{Y}, AA^\top)
$$

Recall that a Gaussian distribution is completely specified by its mean and covariance matrix. We see that the covariance matrix of the joint distribution is

$$
\mathbb{E} \left[ \begin{bmatrix} \tilde{w} \\ \tilde{Y} \end{bmatrix} \begin{bmatrix} \tilde{w}^\top \\ \tilde{Y}^\top \end{bmatrix} \right] = 
\begin{bmatrix} A & B \\ 0 & D \end{bmatrix} \begin{bmatrix} A^\top & 0 \\ B^\top & D^\top \end{bmatrix}
= 
\begin{bmatrix} AA^\top + BB^\top & BD^\top \\ DB^\top & DD^\top \end{bmatrix}
= 
\begin{bmatrix} \Sigma_w & \Sigma_{w,Y} \\ \Sigma_{Y,w} & \Sigma_Y \end{bmatrix}
$$

Matching the corresponding terms, we can express the conditional mean and variance of $\tilde{w} \mid \tilde{Y}$ in terms of these (cross-)covariance matrices:

$$
BD^{-1}\tilde{Y} = B\underbrace{D^\top D^{-T}}_{I} D^{-1}\tilde{Y} = (BD^\top)(DD^\top)^{-1}\tilde{Y} = \Sigma_{w,Y} \Sigma_Y^{-1}\tilde{Y}
$$

$$
AA^\top = AA^\top + BB^\top - BB^\top
= AA^\top + BB^\top - BD^\top D^{-T} D^{-1} DB^\top
= AA^\top + BB^\top - (BD^\top)(DD^\top)^{-1} DB^\top
= \Sigma_w - \Sigma_{w,Y} \Sigma_Y^{-1} \Sigma_{Y,w}
$$
We can then apply the same reasoning to the original setup:

\[
\mathbb{E} \begin{bmatrix} \tilde{Y} \\ \tilde{w}^\top \end{bmatrix} = \begin{bmatrix} R_z R_z^\top + A R_w R_w^\top A^\top & A R_w R_w^\top \\ R_w R_w^\top A^\top & R_w R_w^\top \end{bmatrix} = \begin{bmatrix} \Sigma_Y & \Sigma_{Y,w} \\ \Sigma_{w,Y} & \Sigma_w \end{bmatrix}
\]

Therefore after defining \( \Sigma_z = R_z R_z^\top \), we can read off

\[
\begin{aligned}
\Sigma_w &= R_w R_w^\top \\
\Sigma_Y &= \Sigma_z + A \Sigma_w A^\top \\
\Sigma_{Y,w} &= A \Sigma_w \\
\Sigma_{w,Y} &= \Sigma_w A^\top
\end{aligned}
\]

Plugging this into our estimator yields

\[
\hat{\tilde{w}} = \mathbb{E} [\tilde{w} | \bar{Y} = \bar{y}] \\
= \Sigma_{w,Y} \Sigma_Y^{-1} \bar{y} \\
= \Sigma_w A^\top (\Sigma_z + A \Sigma_w A^\top)^{-1} \bar{y}
\]

One may be concerned because this expression does not take the form we expect – the inverted matrix is hitting \( \bar{y} \) directly, unlike in other solutions we’ve seen. But using the Woodbury matrix identity\(^1\) it turns out that we can rewrite this expression as

\[
\hat{\tilde{w}} = (A^\top \Sigma_z^{-1} A + \Sigma_w^{-1})^{-1} A^\top \Sigma_z^{-1} \bar{y}
\]

which looks more familiar.

\footnotesize
\(^1\) \( (A + UCV)^{-1} = A^{-1} - A^{-1} U (C^{-1} + VA^{-1} U)^{-1} VA^{-1} \)