Hand-in 1 Comments

The OLS point estimate,

$$\hat{\theta}_{OLS} = \underset{\theta}{\operatorname{argmin}}\left\{(\Phi\theta - \mathcal{D})^T(\Phi\theta - \mathcal{D})\right\},$$

corresponds to having $\lambda = 0$, or more precisely, either $\sigma^2 = 0$ or $\sigma_0 = \infty^*$. First, I think it's safe to discard the $\sigma^2 = 0$ case. Sure enough this gives $\lambda = 0$, but having a likelihood with zero variance (a delta function), seems rather un-physical to me as we'll never be able to measure nor model data with infinite accuracy. While the task doesn't explicitly say that $\sigma^2 > 0$, it's specified that data is i.i.d. and that errors are homoscedastic. In other words, we have errors, so $\sigma^2 > 0$.

More importantly, considering $\sigma_0 = \infty$ gives an infinitely wide prior. To me, a normal/Laplace distribution where the standard deviation (or variance) goes to infinity would converge to a uniform distribution on the interval $(-\infty, \infty)$. Using such a prior, i.e. $\mathcal{U}(-\infty, \infty)$, would effectively do nothing since all θ:s have an equal probability. Put differently, if we infer no prior knowledge on θ, the posterior is only affected by the likelihood. Finding the MAP, is, thus, analogous to finding the maximum likelihood. Referring to the lingo in the book (section 2.2.2), having the parameters $\sigma_0 = \infty$ and $\mu = 0$ corresponds to the non-informative limit for the normal prior, and hence the point estimate would be the same as OLS.

To summarize, a powerful aspect of a Bayesian approach for linear regression is that it lets us alter the posterior distribution for model parameters based prior knowledge about θ. In these cases, such knowledge is described by σ_0, and, as illustrated above, an informative prior, $\lambda \neq 0$, truly changes the MAP.

---

* Again,

$$\lambda = \begin{cases} (\sigma/\sigma_0)^2 & \text{w. normal prior} \\ \sigma^2/2\sigma_0 & \text{w. Laplace prior} \end{cases}$$