Hand in 1

Erik Karlsson Öhman

November 12, 2024

To find the maximum a posteriori (MAP) estimator with respect to θ, $\hat{\theta}_{\text{MAP}}$, for an iid data likelihood with homoscedastic errors analyzed with a linear model design matrix $\mathbf{\Phi}$ and parameters $\boldsymbol{\theta}$

$$p(d_i|\boldsymbol{\theta}) = \mathcal{N}(d_i|[\mathbf{\Phi}\boldsymbol{\theta}]_i, \sigma^2) \tag{1}$$

we need to solve

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left[p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] \tag{2}$$

where we may write $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N_p} p(d_i|\boldsymbol{\theta})$ since our data likelihood is iid.

• For the case that we have an uncorrelated normal prior $p(\boldsymbol{\theta}) = \prod_{i=1}^{N_p} \mathcal{N}(\theta_i|0, \sigma_0^2)$ we get

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} \left[\mathcal{N}(\mathcal{D}|\mathbf{\Phi}\boldsymbol{\theta}, \sigma^2\mathbf{1})\mathcal{N}\left(\boldsymbol{\theta}|\mathbf{0}, \sigma_0^2\mathbf{1}\right)\right] \\
&= \arg\max_{\theta} \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_d}{2}} \exp\left\{-\frac{1}{2}\frac{(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})^T(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})}{\sigma^2}\right\}\left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{N_d}{2}} \exp\left\{-\frac{1}{2}\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{\sigma_0^2}\right\}\right] \\
&= \arg\max_{\theta} \left[\left(\frac{1}{2\pi\sigma\sigma_0}\right)^{N_d} \exp\left\{-\frac{1}{2}\frac{(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})^T(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})}{\sigma^2} - \frac{1}{2}\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{\sigma_0^2}\right\}\right] \\
&= \arg\max_{\theta} \left[-\frac{1}{2\sigma^2}\left((\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D})^T(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D}) + \frac{\sigma^2}{\sigma_0^2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right)\right] \\
&= \arg\min_{\theta} \left[\left((\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D})^T(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D}) + \frac{\sigma^2}{\sigma_0^2}\boldsymbol{\theta}^T\boldsymbol{\theta}\right)\right]
\end{aligned} \tag{3}$$

Noting that $\boldsymbol{\theta}^T\boldsymbol{\theta} = \sum_{i=1}^{N_p}\theta_i^2$ and defining $\lambda \equiv \frac{\sigma^2}{\sigma_0^2}$ we arrive at

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \left[(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D})^T(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D}) + \lambda\sum_{i=1}^{N_p}\theta_i^2\right] = \hat{\theta}_{\text{Ridge}} \tag{4}$$

i.e., the ridge estimator.

• For the case of an uncorrelated Laplace parameter prior, $p(\boldsymbol{\theta}) = \prod_{i=1}^{N_p}\mathcal{L}(\theta_i|0,\sigma_0) = \left(\frac{1}{2\sigma_0}\right)^{N_d}\prod_{i=1}^{N_p}\exp\left\{-\frac{|\theta_i|}{\sigma_0}\right\}$ we instead get

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} \left[\mathcal{N}(\mathcal{D}|\mathbf{\Phi}\boldsymbol{\theta}, \sigma^2\mathbf{1})\mathcal{L}\left(\boldsymbol{\theta}|\mathbf{0}, \sigma_0\mathbf{1}\right)\right] \\
&= \arg\max_{\theta} \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_d}{2}} \exp\left\{-\frac{1}{2}\frac{(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})^T(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})}{\sigma^2}\right\}\left(\frac{1}{2\sigma_0}\right)^{N_d} \exp\left\{-\sum_{i=1}^{N_p}\frac{|\theta_i|}{\sigma_0}\right\}\right] \\
&= \arg\max_{\theta} \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N_d}{2}}\left(\frac{1}{2\sigma_0}\right)^{N_d} \exp\left\{-\frac{1}{2}\frac{(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})^T(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})}{\sigma^2} - \sum_{i=1}^{N_p}\frac{|\theta_i|}{\sigma_0}\right\}\right] \\
&= \arg\max_{\theta} \left[-\frac{1}{2\sigma^2}\left((\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta})^T(\mathcal{D}-\mathbf{\Phi}\boldsymbol{\theta}) + 2\frac{\sigma^2}{\sigma_0}\sum_{i=1}^{N_p}|\theta_i|\right)\right] \\
&= \arg\min_{\theta} \left[(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D})^T(\mathbf{\Phi}\boldsymbol{\theta}-\mathcal{D}) + \lambda\sum_{i=1}^{N_p}|\theta_i|\right]
\end{aligned} \tag{5}$$

where $\lambda \equiv 2\frac{\sigma^2}{\sigma_0}$. This is exactly the LASSO regression estimator.

• The quantities $\lambda_{\text{Ridge}} = \frac{\sigma^2}{\sigma_0^2}$ and $\lambda_{\text{LASSO}} = 2\frac{\sigma^2}{\sigma_0}$ are ratios between the width of the likelihood and the width of the prior, and may thus be interpreted as measures of how informative the priors are in relation to the likelihood functions.

• The Ordinary Least Square estimator is given by

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\theta} \Big[(\mathbf{\Phi\theta} - \mathcal{D})^T (\mathbf{\Phi\theta} - \mathcal{D}) \Big] \tag{6}$$

thus if we let $\lambda \to 0$ in the expressions for $\hat{\theta}_{\text{MAP}}$ above we will recover $\hat{\theta}_{\text{OLS}}$. Letting $\lambda \to 0$ is equivalent with letting $\sigma_0^2 \to \infty$. When the variance of the prior goes to infinity we will get a flat prior, i.e. a uniform distribution $p(\boldsymbol{\theta}) = c$ where c is a constant. We may see that setting the prior to $p(\boldsymbol{\theta}) = c$ and calculating gives us

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} \left[\mathcal{N}(\mathcal{D}|\mathbf{\Phi\theta}, \sigma^2\mathbf{1}) \cdot c \right] \\
&= \arg\max_{\theta} \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{N_d}{2}} \exp\left\{ -\frac{1}{2} \frac{(\mathcal{D} - \mathbf{\Phi\theta})^T (\mathcal{D} - \mathbf{\Phi\theta})}{\sigma^2} \right\} \right] \\
&= \arg\min_{\theta} \Big[(\mathbf{\Phi\theta} - \mathcal{D})^T (\mathbf{\Phi\theta} - \mathcal{D}) \Big]
\end{aligned} \tag{7}$$

• Regularization can be interpreted as a way to prohibit overfitting by introducing additional constraints when solving for θ. From a Bayesian perspective this can be interpreted as introducing more information in the form of a prior. In the case when the regularization parameter λ is set to 0 we retrieve the ordinary least squares estimator where we fit our function only to the data, a scenario that may be prone to overfitting. When λ increases so does also the emphasis that is put on the prior when solving for θ, for instance in the case of ridge regression we penalize solutions where the magnitudes of θ_i are not small. This is equivalent to favouring solutions where θ_i is drawn from a normal distribution centered at 0, which makes clear the Bayesian probabilistic connection to regularization. On the other hand the Laplace distribution contains more probability mass in the vicinity of 0 than the normal distribution, something that is reflected in the regularization terms of LASSO regression. The value of $|\theta_i|$ increases more quickly than θ_i^2 near zero so LASSO regression will favour sparse solutions, i.e. where some θ_i are set exactly to zero.