1 The marginal likelihood

We will perform Type-II Maximum-Likelihood estimation so to calculate the Type-II MLE we wish to find the equation for the following marginal:

$$\mathbf{W}' = \arg\max_{\mathbf{W}} \int p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) p(\mathbf{X}) d\mathbf{X}$$

We have the equation for the linear model as:

$$p(\mathbf{Y}, \mathbf{X}, \mathbf{W}) = p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) p(\mathbf{X}) p(\mathbf{W})$$

We assume additive Gaussian noise and a spherical prior so our likelihood and prior are as follows:

$$\begin{aligned} p(y_i | x_i, \mathbf{W}) &= \mathcal{N}(y_i | \mathbf{W} x_i + \mu, \sigma^2 \mathbf{I}) \\ p(x) &= \mathcal{N}(0, \mathbf{I}) \end{aligned}$$

Our marginalisation is the product of two Gaussians which is itself another Gaussian:

$$\begin{aligned} p(y_i | \mathbf{W}) &= \int p(y_i | x_i, \mathbf{W}) p(x_i) \, dx \\ &= \int \mathcal{N}(y_i | \mathbf{W} x_i + \mu, \sigma^2 \mathbf{I}) \, \mathcal{N}(0, \mathbf{I}) \, dx \end{aligned}$$

By replacing these Gaussians with their probability density function and combining we can form a single exponential, by then completing the square with regard to x we can integrate x out to form a new exponent. We re-arrange to find our marginal likelihood:

$$\begin{aligned} p(y_i | \mathbf{W}) &= \mathcal{N}(y_i | \mu, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}) \\ p(\mathbf{Y} | \mathbf{W}) &= \prod_{i=1}^{N} \mathcal{N}(y_i | \mu, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}) \end{aligned}$$

2 The objective function

We use our marginal likelihood from above, we assume the data mean is zero to simplify calculations:

$$p(\mathbf{Y} | \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(y_i | 0, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I})$$

From this we can evaluate, using log rules, our negative log likelihood objective function as follows:

$$
\begin{aligned}
\mathcal{L}(\mathbf{W}) &= -\log\left(\prod_{i=1}^{N}\mathcal{N}(y_i|0,\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})\right)\\
&= -\sum_{i=1}^{N}\log\big(\mathcal{N}(y_i|0,\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})\big)\\
&= -\sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi|\Sigma|}}\right)+\log\left(\exp\Big(-\frac{1}{2}y_i^T\Sigma^{-1}y_i\Big)\right)\\
&= -\sum_{i=1}^{N}\log\frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|^{\frac{1}{2}}}+\log\left(\exp\Big(-\frac{1}{2}y_i^T(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}y_i\Big)\right)\\
&= -\log\big((2\pi)^D|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|\big)^{-\frac{N}{2}}-\sum_{i=1}^{N}-\frac{1}{2}y_i^T(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}y_i\\
&= \frac{N}{2}\left(D\log 2\pi+\log(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|)\right)+\frac{1}{2}\sum_{i=1}^{N}y_i^T(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}y_i
\end{aligned}
$$

From the fact that $\mathrm{Tr}(\mathbf{AB})=\mathrm{Tr}(\mathbf{BA})$ and $\mathrm{Tr}(\mathbf{C}=\mathbf{c})$ for the dimensions in our sum we can evaluate it as follows:

$$
\mathcal{L}(\mathbf{W})=\frac{N}{2}\left(D\log 2\pi+\log(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|)+\mathrm{Tr}\left[\mathbf{Y}(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\right]\right)
$$

3 The derivative of the objective function

To find the gradient of $\mathcal{L}$ we will look at each term in turn, we will make extensive use of [1] to calculate the derivatives of matrices.

$$
\mathcal{L}(\mathbf{W})=\frac{N}{2}\left(\overbrace{\mathrm{Tr}\left[(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{Y}\mathbf{Y}^T\right]}^{A}+\overbrace{\log(|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|)}^{B}+\overbrace{D\log 2\pi}^{C}\right)
$$

C is a constant term so we shall discard that. To find the derivative of B we use the rules that $\partial(\log(\det(\mathbf{X})))=\mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X})$ and that $\sigma^2\mathbf{I}$ is constant with respect to $\mathbf{W}$ and so evaluate B to:

$$
\frac{\partial B}{\partial\mathbf{W}}=\mathrm{Tr}\left((\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\partial(\mathbf{W}\mathbf{W}^T)\right)
$$

We also have the rules that $\partial(\mathbf{XY})=(\partial\mathbf{X})\mathbf{Y}+\mathbf{X}(\partial\mathbf{Y})$ and that $\frac{\partial\mathbf{X}}{\partial\mathbf{X}_{ij}}=\mathbf{J}^{ij}$ where $\mathbf{J}$ is the single-entry matrix, having 1 at (i,j) and 0 elsewhere and so can reduce B to:

$$
\frac{\partial B}{\partial\mathbf{W}_{ij}}=\mathrm{Tr}\left((\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}(\mathbf{J}^{ij}\mathbf{W}^T+\mathbf{W}\mathbf{J}^{ijT})\right)
$$

Using the rule that $\partial(\mathrm{Tr}(\mathbf{X})) = \mathrm{Tr}(\partial\mathbf{X})$, the chain rule and that $\frac{\partial \mathbf{X}\mathbf{C}\mathbf{X}^T}{\partial \mathbf{C}} = \mathbf{X}^T\mathbf{X}$ we can find the derivative of A:

$$\begin{aligned}\frac{\partial A}{\partial \mathbf{W}} &= \mathrm{Tr}\left(\partial\Big(\mathbf{Y}(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\Big)\right)\\ &= \mathrm{Tr}\left(\mathbf{Y}^T\mathbf{Y}\partial\Big((\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\Big)\right)\end{aligned}$$

As $\partial(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1}$ we can evaluate A to:

$$\frac{\partial A}{\partial \mathbf{W}_{ij}} = \mathrm{Tr}\left(\mathbf{Y}^T\mathbf{Y}\Big(-(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{J}^{ij}\mathbf{W}^T + \mathbf{W}\mathbf{J}^{ijT})(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\Big)\right)$$

And so by combining our terms A and B we have the gradient for our log likelihood $\mathcal{L}$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ij}} = \frac{N}{2}\Bigg(&\mathrm{Tr}\Big[\mathbf{Y}^T\mathbf{Y}\Big(-(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{J}^{ij}\mathbf{W}^T + \mathbf{W}\mathbf{J}^{ijT})(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\Big)\Big]\\ &+ \mathrm{Tr}\Big[(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{J}^{ij}\mathbf{W}^T + \mathbf{W}\mathbf{J}^{ijT})\Big]\Bigg)\end{aligned}$$

References

[1] K. B. Petersen and M. S. Petersen. *The Matrix Cookbook.* November 2012.