Question 19

Derivation of the Gradient

Carl Henrik Ek

October 19, 2017

In question 19 we will derive the gradient for Principled Component Analysis, in the formulation in the book the author derive a closed form solution for the Type-II Maximum likelihood, this is rather tricky so insted we will do a gradient based approach. Deriving these are still not a walk in the park if you are not comfortable with matrix derivatives. I will here go through a couple of tips on how to get you there.

The first thing that we need to do is to write up the objective function, this is going to consist of three terms which looks like this,

$$\mathcal{L}(\mathbf{W}) = \text{constant} + \log|\mathbf{C}(\mathbf{W})| + \sum_{i}^{N} \mathbf{y}_i^{\mathrm{T}}(\mathbf{C}(\mathbf{W}))^{-1}\mathbf{y}_i$$

Now we need to compute the derivatives of $\mathbf{C}(\mathbf{W})$ with respect to $\mathbf{W}$. To do this lets first rewrite everything on matrix form so that we can remove the sum from the objective function. First note that,

$$\sum_i \mathbf{x}_i\mathbf{x}_i = \text{tr}\left(\begin{bmatrix} \leftarrow & \mathbf{x}_1 & \rightarrow \\ \leftarrow & \mathbf{x}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_N & \rightarrow \end{bmatrix}\begin{bmatrix} \leftarrow & \mathbf{x}_1 & \rightarrow \\ \leftarrow & \mathbf{x}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_N & \rightarrow \end{bmatrix}^{\mathrm{T}}\right),$$

this means that we can rewrite the objective function as,

$$\mathcal{L}(\mathbf{W}) = \text{constant} + \log|\mathbf{C}(\mathbf{W})| + \text{tr}\left(\mathbf{Y}(\mathbf{C}(\mathbf{W}))^{-1}\mathbf{Y}^{\mathrm{T}}\right).$$

Now we have two terms we need to take derivatives of, both of these include the same matrix $\mathbf{C}$ in one form of the other so it will come in handy to take this derivative first,

$$\frac{\partial \mathbf{C}}{\partial \mathbf{W}_{ij}} = \frac{\partial \mathbf{W}\mathbf{W}^{\mathrm{T}}}{\partial \mathbf{W}_{ij}}.$$

We will now use the excellent Matrix Cookbook [1] to find the rules that we need to figure this one out. If you use the version from 2012 URL we can first use Eq. 37 to rewrite the the product,

$$\partial(\mathbf{X}\mathbf{Y}) = (\partial\mathbf{X})\mathbf{Y} + \mathbf{X}(\partial\mathbf{Y}).$$

We can then combine this with Eq. 32 which states,

$$\frac{\partial \mathbf{X}_{kl}}{\partial \mathbf{X}_{ij}} = \delta_{ik}\delta_{lj},$$

where δ_{ij} is the kronecker delta function,

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

This leads to the following derivative,

$$\frac{\partial \mathbf{W}\mathbf{W}^{\mathrm{T}}}{\partial \mathbf{W}_{ij}} = \mathbf{W}\frac{\partial \mathbf{W}^{\mathrm{T}}}{\partial \mathbf{W}_{ij}} + \frac{\partial \mathbf{W}}{\partial \mathbf{W}_{ij}}\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{J}_{ij} + \mathbf{J}_{ji}\mathbf{W}^{\mathrm{T}}$$

where we $\mathbf{J}_{ij}$ is a matrix who has all zero entries except for $(\mathbf{J}_{ij})_{ij} = 1$.

Now lets strart by tackling the first term, the derivative of the log determinant.

$$\frac{\partial}{\partial \mathbf{W}_{ij}} \log|\mathbf{C}|$$

We will start by using Eq. 43 that states that,

$$\partial \log|\mathbf{X}| = \mathrm{tr}\left(\mathbf{X}^{-1}\partial \mathbf{X}\right).$$

We can now rewrite this as,

$$\frac{\partial}{\partial \mathbf{W}_{ij}} \log|\mathbf{C}| = \mathrm{tr}\left(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \mathbf{W}_{ij}}\right).$$

So now we have our first term and we can move on to the second. This term is a derivative of a trace we can now use Eq. 36 which states,

$$\partial\left(\mathrm{tr}(\mathbf{X})\right) = \mathrm{tr}\left(\partial \mathbf{X}\right).$$

This means that our second term becomes,

$$\frac{\partial}{\partial \mathbf{W}_{ij}} \mathrm{tr}\left(\mathbf{Y}(\mathbf{C})^{-1}\mathbf{Y}^{\mathrm{T}}\right) = \mathrm{tr}\left(\frac{\partial}{\partial \mathbf{W}_{ij}}\mathbf{Y}(\mathbf{C})^{-1}\mathbf{Y}^{\mathrm{T}}\right).$$

Now we want to break the quadratic form and we will do this by using the chain-rule to do the derivative,

$$\mathrm{tr}\left(\frac{\partial}{\partial \mathbf{W}_{ij}}\mathbf{Y}(\mathbf{C})^{-1}\mathbf{Y}^{\mathrm{T}}\right) = \mathrm{tr}\left(\frac{\partial}{\partial \mathbf{C}}(\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^{\mathrm{T}})\frac{\partial \mathbf{C}^{-1}}{\partial \mathbf{W}_{ij}}\right) = \mathrm{tr}\left((\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\frac{\partial \mathbf{C}^{-1}}{\partial \mathbf{W}_{ij}}\right).$$

Now we have isolated the derivative and we are nearly there. The last rule we will use is to use the derivative of a matrix inverse Eq. 40,

$$\partial \mathbf{X}^{-1} = -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1}.$$

This means we can reach the derivative of the last term as,

$$\mathrm{tr}\left((\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\frac{\partial \mathbf{C}^{-1}}{\partial \mathbf{W}_{ij}}\right) = \mathrm{tr}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\left(-\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \mathbf{W}_{ij}}\mathbf{C}^{-1}\right)\right),$$

where we know the derivative of the inner-most term as we computed it first.

So now we have the full derivative, as we can see the dimensionality makes sense, we take the derivative of our objective function which is a scalar with another scalar and therefore expect a scalar back, as both of our terms are traces of matrices this makes sense. Now you just have to code this up and put this into the optimiser and see what solution comes out.

References

[1] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, November 2012. Version 20121115.