3.c

To find the maximum likelihood solution W_{ML}, we need to take the derivative w.r.t W of the log-likelihood function computed above and equal it to zero. We start by computing the derivative w.r.t W^T and proceed to prove that its equal to the transpose of the derivative w.r.t W. So we have:

$$\frac{\partial}{\partial W^T}(-\frac{1}{2}(NK\log 2\pi + N\log\det\Sigma + \sum_{i=1}^{N}(t_i - y(x_i, W))^T\Sigma^{-1}(t_i - y(x_i, W)))) =$$

We remove the first two terms $\frac{NK}{2}\log 2\pi$ and $\frac{N}{2}\log\det\Sigma$ since their derivative w.r.t W^T is zero. So we get:

$$= \frac{\partial}{\partial W^T}(-\frac{1}{2}(\sum_{i=1}^{N}(t_i - y(x_i, W))^T\Sigma^{-1}(t_i - y(x_i, W)))) =$$

And now by the sum rule:

$$= -\frac{1}{2}(\sum_{i=1}^{N}\frac{\partial}{\partial W^T}(t_i - y(x_i, W))^T\Sigma^{-1}(t_i - y(x_i, W))) = -\frac{1}{2}(\sum_{i=1}^{N}\frac{\partial}{\partial W^T}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i))) =$$

Making use of the hint given to us $\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T$ and the fact that the derivative w.r.t the transpose is equal to the transpose of the derivative w.r.t the orignal matrix , we have:

$$= -\frac{1}{2}(\sum_{i=1}^{N} -2(\Sigma^{-1}(t_i - W^T\phi(x_i))\phi(x_i)^T)^T = \sum_{i=1}^{N}\phi(x_i)(t_i^T - \phi(x_i)^T W)\Sigma^{-1} =$$

Transforming it to the matrix form, we get:

$$= \Phi^T (T - \Phi W)\Sigma^{-1} = \Phi^T T\Sigma^{-1} - \Phi^T \Phi W \Sigma^{-1}$$

Now equaling it to zero, we have:

$$\Phi^T T\Sigma^{-1} - \Phi^T \Phi W \Sigma^{-1} = 0$$

$$\Phi^T \Phi W \Sigma^{-1} = \Phi^T T \Sigma^{-1}$$

But we know that the covariance matrix is symmetric and positive definite, thus it has an inverse - that we've been using - and we can do:

$$\Phi^T \Phi W \Sigma^{-1}\Sigma = \Phi^T T \Sigma^{-1}\Sigma$$

$$\Phi^T \Phi W = \Phi^T T$$

We also know that $\Phi^T\Phi$ should be invertible as long as we get linear independent x input vectors. Assuming we do, we then have:

$$W_{ML} = (\Phi^T \Phi)^{-1}\Phi^T T$$

So now we have shown that the maximum likelihood solution is $W_{ML} = (\Phi^T\Phi)^{-1}\Phi^T T$ and thus is independent of the covariance matrix Σ. We just need to show that this result is valid by proving the derivative of the likelihood w.r.t W^T is equal to the transpose of the derivative of the likelihood w.r.t W:

$$-\frac{1}{2}(\sum_{i=1}^{N}\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i)))$$

Focusing on the derivative term and expanding to then use index notation, we have:

$$\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i)) = \frac{\partial}{\partial W}(t_i^T - \phi(x_i)^T W)\Sigma^{-1}(t_i - W^T\phi(x_i)) =$$

$$= \frac{\partial}{\partial W}(t_i^T\Sigma^{-1} - \phi(x_i)^T W\Sigma^{-1})(t_i - W^T\phi(x_i)) =$$

$$= \frac{\partial}{\partial W}(t_i^T\Sigma^{-1}t_i - \phi(x_i)^T W\Sigma^{-1}t_i - t_i^T\Sigma^{-1}W^T\phi(x_i) + \phi(x_i)^T W\Sigma^{-1}W^T\phi(x_i))$$

By the sum rule, we can compute each derivative separately:

1.

$$\frac{\partial}{\partial W}t_i^T\Sigma^{-1}t_i = 0$$

2.

$$\frac{\partial}{\partial W}\phi(x_i)^T W\Sigma^{-1}t_i \implies \sum_{p=1}^{M}\sum_{q=1}^{K}\sum_{r=1}^{K}\frac{\partial}{\partial W_{m,n}}\phi(x_i)_p W_{p,q}\Sigma^{-1}_{q,r}(t_i)_r = \sum_{r=1}^{K}\phi(x_i)_m\Sigma^{-1}_{n,r}(t_i)_r \implies$$

$$\implies \frac{\partial}{\partial W}\phi(x_i)^T W\Sigma^{-1}t_i = \phi(x_i)(\Sigma^{-1}t_i)^T$$

3.

$$\frac{\partial}{\partial W}t_i^T\Sigma^{-1}W^T\phi(x_i) \implies \sum_{p=1}^{K}\sum_{q=1}^{K}\sum_{r=1}^{M}\frac{\partial}{\partial W_{m,n}}(t_i)_p\Sigma^{-1}_{p,q}W_{q,r}\phi(x_i)_r = \sum_{p=1}^{K}(t_i)_p\Sigma^{-1}_{p,m}\phi(x_i)_n \implies$$

$$\implies \frac{\partial}{\partial W} t_i^T \Sigma^{-1} W^T \phi(x_i) = (t_i^T \Sigma^{-1})^T \phi(x_i)^T = \phi(x_i)(\Sigma^{-1} t_i)^T$$

4.

$$\frac{\partial}{\partial W} \phi(x_i)^T W \Sigma^{-1} W^T \phi(x_i) \implies \sum_{p=i}^{M} \sum_{q=1}^{K} \sum_{r=1}^{K} \sum_{s=1}^{M} \frac{\partial}{\partial W_{m,n}} \phi(x_i)_p W_{p,q} \Sigma^{-1}_{q,r} W_{r,s} \phi(x_i)_s =$$

By the product rule:

$$= \sum_{p=i}^{M} \sum_{q=1}^{K} \sum_{r=1}^{K} \sum_{s=1}^{M} (\phi(x_i)_m \Sigma^{-1}_{n,r} W_{r,s} \phi(x_i)_s + \phi(x_i)_p W_{p,q} \Sigma^{-1}_{q,m} \phi(x_s)_n) =$$

$$= \sum_{r=1}^{K} \sum_{s=1}^{M} \phi(x_i)_m \Sigma^{-1}_{n,r} W_{r,s} \phi(x_i)_s + \sum_{p=i}^{M} \sum_{q=1}^{K} \phi(x_i)_p W_{p,q} \Sigma^{-1}_{q,m} \phi(x_s)_n \implies$$

$$\implies \frac{\partial}{\partial W} \phi(x_i)^T W \Sigma^{-1} W^T \phi(x_i) = \phi(x_i)(\Sigma^{-1} W^T \phi(x_i))^T + \phi(x_i)\phi(x_i)^T W \Sigma^{-1} =$$

Using the symmetry of Σ and thus Σ^{-1} :

$$= \phi(x_i)\phi(x_i)^T W\Sigma^{-1} + \phi(x_i)\phi(x_i)^T W\Sigma^{-1} = 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1}$$

Putting all the terms together, we have:

$$\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i)) = -\phi(x_i)(\Sigma^{-1}t_i)^T - \phi(x_i)(\Sigma^{-1}t_i)^T + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} =$$

$$= -2\phi(x_i)(\Sigma^{-1}t_i)^T + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} = -2\phi(x_i)t_i^T\Sigma^{-1} + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} =$$
$$= -2\phi(x_i)(t_i^T - \phi(x_i)^T W)\Sigma^{-1}$$

And if we put the obtained equation back to the original expression, we get:

$$-\frac{1}{2}(\sum_{i=1}^{N}\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i))) = \sum_{i=1}^{N}\phi(x_i)(t_i^T - \phi(x_i)^T W)\Sigma^{-1}$$

Exactly the expression we got by using the hint and thus proving that indeed the derivative of the likelihood w.r.t W^T is equal to the transpose of the derivative of the likelihood w.r.t W.

So we can finally state that $W_{ML} = (\Phi^T\Phi)^{-1}\Phi^T T$.