Assignment 1 - Report

Pietro Alovisi

11-17-2018

Question 1

The gaussian function is a unimodal distribution, which means that has only one mode and for this particular distribution it coincides with the mean. So in this case we are assuming that value of the deterministic function f for a given $\mathbf{x}$ is the mean value of the distribution of the target. This can be rephrased as assuming a determinsitic model $f(\mathbf{x})$ that generates realizations with a random error ε that distributes as $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Putting everyting togheter we get:

$$\mathbf{t} = f(\mathbf{x}) + \varepsilon$$

A prior oservation about the covariance is that we are assuming homoscedasticity, that is the variance of $\mathbf{t}$ is not dependent on the input vector $\mathbf{x}$.

The spherical covariance matrix means implies two facts:

- All the scalar random variables t_j of the vector $\mathbf{t_i}$ have the same variance σ^2.
- The fact that the covariance matrix is diagonal means that all the output sclar component t_j of the vector $\mathbf{t_i}$ are independent one another.

Question 2

If we do not assume independence of the samples, we must turn to the joint probability distribution

$$p(\mathbf{T}|f, \mathbf{X}) = p(\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_N}|f, \mathbf{X})$$

Question 3

Equation 5 is a linear transformation of a normal distribution which, from its properties, is again a normal distribution equal to:

$$p(\mathbf{t_i}) \sim \mathcal{N}(\mathbf{W}\,\mathbf{x_i}, \sigma^2\mathbf{I})$$

Still assuming conditionally independent samples, from 3 the likelyhood is just:

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{t_i}|\mathbf{W}\,\mathbf{x_i}, \sigma^2\mathbf{I})$$

Which we can also write by vectorising the whole, by noting that since all the $\mathbf{t_i}$ have the same variance, the exponents in the probability density function sum up.

$$\begin{aligned} p(\mathbf{T}|\mathbf{X}, \mathbf{W}) &= \mathcal{N}(\mathbf{X}\mathbf{W}^T, \mathbf{I}, \sigma^2\mathbf{I}) = \\ &= \frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})} = \\ &= \frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}Tr\left((\mathbf{X}\mathbf{W}^\mathbf{T}-\mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T}-\mathbf{T})^T\right)} \end{aligned}$$

Where we substituted the expression at the exponent $\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$ with $Tr\big((\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})^T\big)$ by noting that the summation is just the sum of the diagonal of the matrix $(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})(\mathbf{X}\mathbf{W}^\mathbf{T} - \mathbf{T})^T$.

Question 4

The two penalization terms can be obtained from the prior. First let's do the one for the L_2 norm, and then we will generalize to the L_1. We can write the prior on W as:

$$p(W) = \frac{1}{\tau^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{tr((W-W_0)(W-W_0)^T)}{2\tau^2}} = \frac{1}{\tau^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{\sum_i^N w_i^T \cdot w_i}{2\tau^2}}$$

If we multiply with the expression computed above for the lykelyhood $p(\mathbf{T}|\mathbf{X}, \mathbf{W})$ we get:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{T}) \propto e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i}) - \frac{1}{2\tau^2}\sum_i^N w_i^T \cdot w_i}$$

Where w have disregarded the multiplicatove factor in front of the exponent, because by taking the log will lead to a costant factor. Now we take the negative logarithm:

$$-log(p(\mathbf{W}|\mathbf{X},\mathbf{T})) \propto \frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i}) + \frac{1}{2\tau^2}\sum_i^N w_i^T \cdot w_i$$

Where we can easily see the penalizing factor:

$$\frac{1}{2\tau^2}\sum_i^N w_i^T \cdot w_i = \frac{1}{2\tau^2}\sum_i^N \|w_i\|_{L_2}$$

Of course the proper extension to the L_1 norm will lead to the penalizing term:

$$\frac{1}{2\tau^2}\sum_i^N |w_i| = \frac{1}{2\tau^2}\sum_i^N \|w_i\|_{L_1}$$

Using L_1 norm will perform some kind of dimensionality reduction by setting some variables to 0, while the quadratic term will try to balanced the parameter. We can see this effect by inspecting the derivative of the penalizing term: for the $L1$ norm it is always constant, while for the $L2$ it decreases as we get closer to zero, this means that in $L2$ optimizing values that are close to the origin does not get me any decrease in the penalazing term, while if I take a value far away from the origin then this will decrease a lot my penalizing term.

We can aslo see visually by looking at the iso-contours of these functions in figure 1.

We can see that the corners of the square lie on the axis, so where one of the two variables is zero.

$w^T w$: for the L_2 norm which is just the Froebenius norm $|w|_F$

These two priors will introduce an additive term depending on the model parameter $|w|$, which would be of second order for the L_2 metric, and a first order L_1 for the other one.

Question 5

We will use the square completion to perform this task. So we assume that the output is normal with the following parametrs:

$$p(W) = \frac{1}{\xi}\cdot e^{-tr((W-W_0)\Sigma^{-1}(W-W_0)^T)} =$$

$$= \frac{1}{\xi}\cdot e^{-tr(W\Sigma^{-1}W^T)} e^{2\cdot tr(W\Sigma^{-1}W_0^T)} e^{-tr(W_0\Sigma^{-1}W_0^T)}$$

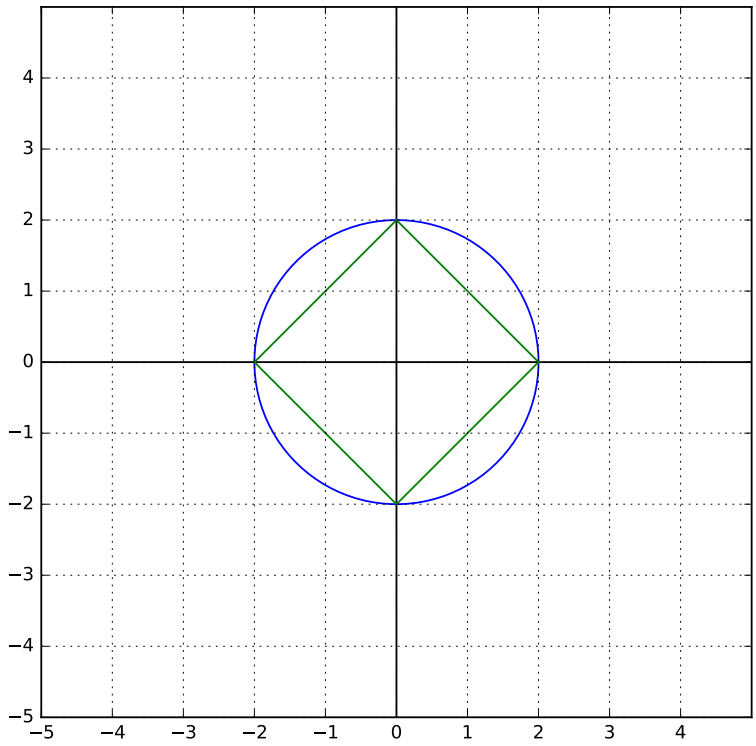


Figure 1: Showing L1 and L2 differences.

Where ξ is just the normlaizing factor to make the integral of the function 1.

Now we will take the product of the prior over W, and the likelyhood $p(\mathbf{t_i}|\mathbf{x_i}, \mathbf{W})$.

$$p(\mathbf{t_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$

Since they have the same dimensions, we can do merge the two traces:

$$p(\mathbf{W}|\mathbf{t_i}, \mathbf{x_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr(\mathbf{t_i}\mathbf{t_i}^T)} e^{\frac{1}{\sigma^2}tr(W\mathbf{x_i}\mathbf{t_i}^T)} e^{-\frac{1}{2\sigma^2}tr(W\mathbf{x_i}\mathbf{x_i}^TW^T)} e^{-\frac{1}{2\tau^2}tr(WW^T)} e^{\frac{1}{\tau^2}tr(WW_0^T)} e^{-\frac{1}{2\tau^2}tr(W_0W_0^T)} =$$
$$= e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{x_i}\mathbf{x_i}^T)W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

Then assuming the independence of the t_i we can get to the full posterior.

$$p(\mathbf{W}|\mathbf{T}, \mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\sum_i \mathbf{x_i}\mathbf{x_i}^T)W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

$$p(\mathbf{W}|\mathbf{T}, \mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X})W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T}+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T}+\frac{1}{2\tau^2}W_0W_0^T)}$$

Where we substituted $\sum_i \mathbf{x_i}\mathbf{t_i}^T$ with $\mathbf{X}^T\mathbf{T}$. This can be demostrated to be true, and so I do in appendix A.

Now we can retrieve the variance and the mean of our prior.

Question 6

To comment the prior we will analize its two components. The least important is the mean, which is set arbitrarely to 0, which means that the functions we'll have zero mean. The most important component is the covariance, that is computed as a kernel function. The kernel function should implement some kind of "closeness measure" between two points x_i and x_j, with the kernel having high values id x_i is similar to x_j, low otherwise. This function sets the correlation between two points, so if x_i and x_j are close, their values will be high correlated, on the opposite side if $k(x_i, x_j) = 0$, then the two values y_i and y_j are independent (works only assuming the distribution normal). Basically the covariance function defines a transfer of information between one point and the other.

Put figure(s) here

Question 7

If we also assume that $\mathbf{X}$ and θ are random variables, we can easily decompose the formula into:

$$p(\mathbf{T}, \mathbf{X}, \mathbf{f}, \theta) = p(\mathbf{T}, \mathbf{f}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)$$

Moreover it's safe to assume that $\mathbf{X}$ and θ are independent, which lets me factor even more the formula into:

$$p(\mathbf{T}, \mathbf{f}|\mathbf{X}, \theta)p(\mathbf{X}, \theta) = p(\mathbf{T}, \mathbf{f}|\mathbf{X}, \theta)p(\mathbf{X})p(\theta)$$

I can use the chain rule one again on the first term to get:

$$p(\mathbf{T}, \mathbf{f}|\mathbf{X}, \theta)p(\mathbf{X})p(\theta) = p(\mathbf{T}|\mathbf{f}, \mathbf{X}, \theta)p(\mathbf{f}|\mathbf{X}, \theta)p(\mathbf{X})p(\mathbf{X})p(\theta)$$

Where we know that $p(\mathbf{f}|\mathbf{X}, \theta)$ is a multivariate normal distribution for the definition of the gaussian processes. While the term $p(\mathbf{T}|\mathbf{f}, \mathbf{X}, \theta)$ can be further expanded using the relation between t_i and f_i which, being conditioned is known.

$$p(\mathbf{t_i} = t^*|\mathbf{f} = f^*, \mathbf{X}, \theta) = p(f^* + \varepsilon = t^*|\mathbf{f} = f^*, \mathbf{X}, \theta) = p(\varepsilon = t^* - f^*|\mathbf{f} = f^*, \mathbf{X}, \theta)$$

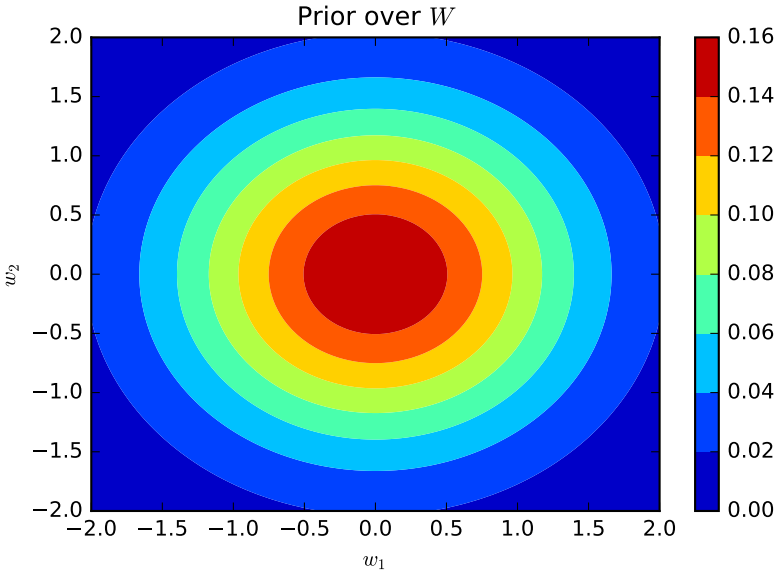
So it distributes just like the error ε.

Question 8

$$p(\mathbf{T}|\mathbf{X}, \theta) = \int p(\mathbf{T}|\mathbf{f}, \mathbf{X}, \theta)p(\mathbf{f}|\mathbf{X}, \theta)df$$

We still condition in θ because we assumed it as a constant, it could be marginalized if we have had assumed it was a random variable. In this form is useful for hyperparameter optimization.

Question 9



Which varianc eof the prior should I choose?

Figure 2: Posterior over the parameters after observing one point

Question 10

The lenghtscale defines a "unit of measure" between the two points. Since it divides the difference between two points, if the lenghtscale is low the two points will be less correlated, if the value of the lenghtscale is high, then they will be highly correlated.

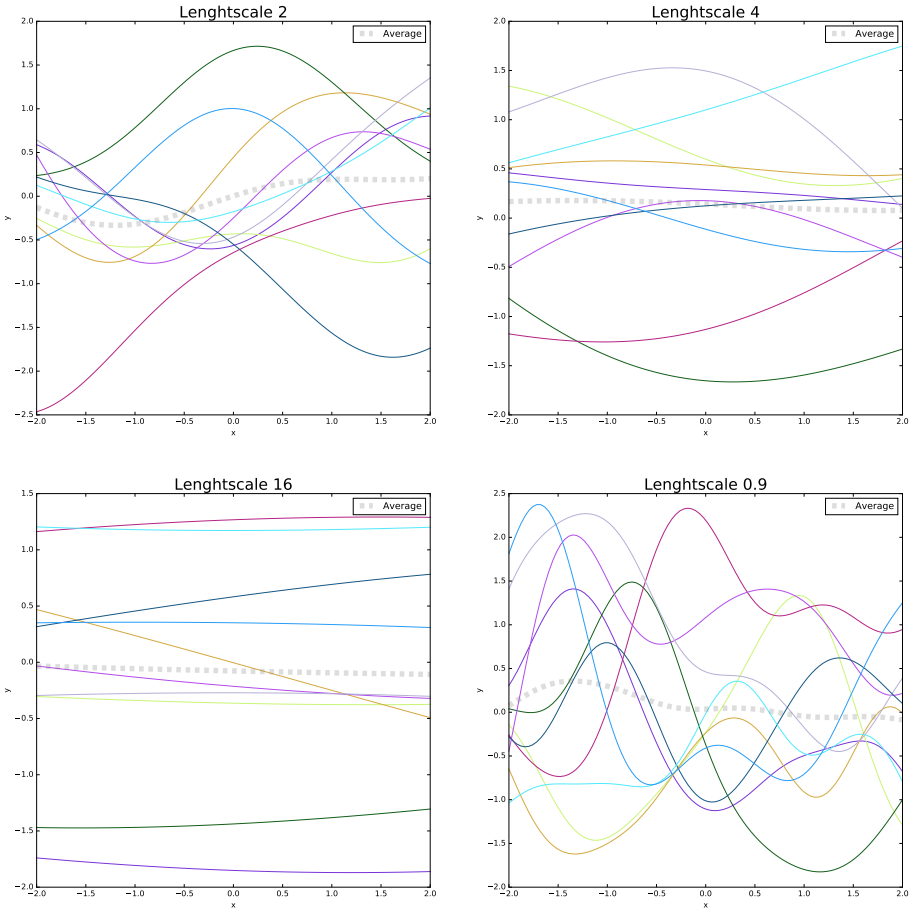


Figure 3: Samples from a gaussian process

Question 11

Question 12

The preference is that our variable X is a normal distribution whose elemetns are independent, and distribute around zero.

Question 13

Since the model is:

$$\mathbf{y_i} = \mathbf{W}\mathbf{x_i} + \varepsilon$$

The mean of each $\mathbf{y_i}$ is :

$$\mathbb{E}(\mathbf{y_i}) = \mathbb{E}(\mathbf{W}\mathbf{x_i} + \varepsilon)$$
$$\mathbb{E}_X(\mathbf{y_i}) = \mathbf{W}\,\mathbb{E}_X(\mathbf{x_i}) + \mathbb{E}_X(\varepsilon)$$
$$\mathbb{E}_X(\mathbf{y_i}) = 0 + 0$$

We can describe the probability by only the first 2 moments of the random variable:

$$\mathrm{Var}(\mathbf{y_i}) = \mathrm{Var}(\mathbf{W}\mathbf{x_i} + \varepsilon)$$

Since they are uncorrelated, we can write:

$$\mathrm{Var}(\mathbf{y_i}) = \mathrm{Var}(\mathbf{W}\mathbf{x_i}) + \mathrm{Var}(\varepsilon)$$
$$\mathrm{Var}(\mathbf{y_i}) = \mathbf{W}\mathrm{Var}(\mathbf{x_i})\mathbf{W}^T + \sigma^2\mathbf{I} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Since each y_i is independent with each other, we can combine the results we got into the distribution:

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

$$p(\mathbf{Y}|\mathbf{W}) \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}, \mathbf{I})$$

Question 14

MLE

From the derived distribution in question 3 we can compute the log likelihood:

$$log(p(\mathbf{Y}|\mathbf{X}, \mathbf{W})) = log\left(\frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i}-\mathbf{t_i})^T(\mathbf{W}\mathbf{x_i}-\mathbf{t_i})}\right) =$$

$$= -log\left(\sigma^2(2\pi)^{\frac{D}{2}}\right) - \frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$$

In the maximization we disregard the constant factor $-log\left(\sigma^2(2\pi)^{\frac{D}{2}}\right)$,and then remove also the multiplicative constant in the second term $\frac{1}{2\sigma^2}$. So we are left with the maximization of:

$$\arg\max_W -\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$$

Which is clealy the generalization of the sum of residual square for vectorial outputs.

MAP

We can derive the expression starting from the previous part of the question.

$$log\left(p(\mathbf{Y}|\mathbf{X},\mathbf{W})\cdot p(\mathbf{W})\right) = log(p(\mathbf{Y}|\mathbf{X},\mathbf{W})) + log(p(\mathbf{W}))$$

The first term of the summation is the MLE term from before, while the second one I have already computed in qustion 4 as:

$$log(p(\mathbf{W})) = -log\left(\tau^2(2\pi)^{\frac{D}{2}}\right) - \frac{1}{2\tau^2}\sum_i^N (\mathbf{w_i})^T(\mathbf{w_i})$$

Again we can disregard the constant term at the begining, but we need to keep the multiplicative terms both for the prior and for the least square. Putting everything together we get:

$$\arg\max_W \left\{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i}) - \frac{1}{2\tau^2}\sum_i^N (\mathbf{w_i})^T(\mathbf{w_i})\right\}$$

Where the second term acts as a reguralizing term.

Type II ML

$$log\left(\int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X}\right) = log\left(p(\mathbf{Y}|\mathbf{W})\right) =$$

$$log\left(\prod_{i=0}^N p(\mathbf{y_i}|\mathbf{x_i},\mathbf{W})\right) =$$

$$= \sum_{i=0}^{N} log\left(p(\mathbf{y_i}|\mathbf{x_i}, \mathbf{W})\right) =$$

If we substitute with the expression for $p(\mathbf{y_i}|\mathbf{x_i}, \mathbf{W})$ we still have a log of a normal distribution.

$$= -\sum_{i=0}^{N} log\left((det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})2\pi^D)^{\frac{1}{2}}\right) - \sum_{i=0}^{N} \mathbf{y_i}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y_i}$$

The two expression in equation 25 are equal because the denominator (the evidence) is constant for any choice of the model parameter W. The evidence only changes if we choose another model.

Type-II Maximum-Likelihood is a sensible way of learning the parameters because we first use the bayesian approach to avoid the overfit on data, and then we maximize the hyperparameter, which cannot overfit, because it is not backed by data.

Question 15

Question 16

Question 17

This is the simples model because it is uninformative, each data set is equally likely.

Question 18

Question 19

Question 20

Question 21

Question 22

Question 23

Question 24

Appenddix A

Demonstration of $\sum_i \mathbf{x_i}\mathbf{x_i}^T = \mathbf{X}^T\mathbf{X}$

Suppose we have a matrix X:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{x_N}^T \end{bmatrix}$$

We can decompose this matrix as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T \end{bmatrix}$$

From the previous it immediatly follows that its transpose can be expressed as:

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{x_2} & \dots & \mathbf{0} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n} \end{bmatrix}$$

Now if we multiply the decomposed version of the matrix $\mathbf{X}^T$ together with $\mathbf{X}$ and apply the distributive property we get:

$$\mathbf{X}^T\mathbf{X} = \left(\begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n} \end{bmatrix}\right) \cdot \left(\begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T \end{bmatrix} \right)$$

From which we can see that the multiplication of any corresponding matrices containing the same vector $\mathbf{x_i}$ we get:

$$\begin{bmatrix}\mathbf{0} & \dots & \mathbf{x_i} & \dots & \mathbf{0}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} = \mathbf{x_i} \cdot \mathbf{x_i}^T$$

While by multipling two matrices containing different vectors we get:

$$\begin{bmatrix}\mathbf{0} & \dots & \mathbf{x_j} & \dots & \mathbf{0}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T\end{bmatrix} = \mathbf{0}$$

We can therfore conclude that:

$$\mathbf{X}^T\mathbf{X} = \sum_i \mathbf{x_i}\mathbf{x_i}^T$$

If we have 2 different matrices $\mathbf{X}$ and $\mathbf{Y}$ we can repeat the procedure and conclude that: