
Coursework 1 - Modelling

Jonas Osborn
jo14944

Tristan Saunders
ts16802

Corin Varney
cv14985

1 The prior

1.1 Theory

Question 1 1. Choosing a gaussian is a sensible thing to do because it makes calculation much easier, Gaussians are quite easily multiplied and integrated which are things you often have to do in machine learning contexts e.g. multiplying to compute a joint distribution or integrating to marginalise out a variable. Choosing a Gaussian likelihood also fits with our model because we often assume errors observation errors to be independent and identically distributed which due to the Central Limit Theorem (which states the distribution of the sum of a large enough number of independent, identically distributed variables will be approximately normally distributed) means our model ends up being a Gaussian.

2. Choosing a spherical covariance matrix for the likelihood means that we are assuming the different dimensions of y_i to be independent and identically distributed. As they are independent they do not covary with each other and so the covariance matrix is diagonal. As they are identically distributed they all have the same variance. Therefore the covariance matrix is spherical.

Question 2

$$\begin{aligned} p(\mathbf{Y}|f, \mathbf{X}) &= p(y_1, \dots, y_{N-1}, y_N | f, \mathbf{X}) \\ &= p(y_N | y_{N-1}, \dots, y_1, f, \mathbf{X}) p(y_{N-1} | y_{N-2}, \dots, y_1, f, \mathbf{X}) \dots p(y_1 | f, \mathbf{X}) \end{aligned}$$

1.1.1 Linear regression

Question 3 We assume that the data is independent and so the joint probability is a product.

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) &= \prod_i^N p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{W}) \\ &= \prod_i^N \mathcal{N}(\mathbf{W} \mathbf{x}_i, \sigma^2 \mathbf{I}) \end{aligned}$$

Question 4 A conjugate prior is one that is conjugate to the posterior, meaning they are in the same family of distributions. A conjugate prior is useful as it gives a closed-form solution for the posterior. If we didn't choose a conjugate prior then numerical integration may be necessary to calculate the posterior which may mean the solution is potentially intractable. The conjugate prior for a Gaussian posterior is a Gaussian.

Question 5 Just as encoding the preference in a L_2 norm is equivalent to having a Gaussian prior, encoding the preference using a L_1 norm is equivalent to having a Laplace prior. This is because the Laplace distribution estimates median rather than the mean estimated by the Gaussian and median minimises the L_1 norm and mean the L_2

The shape of the laplace distribution's probability density function, with it's higher peak around zero compared to the probability density function of a Gaussian means that more co-efficients are likely to be equal to zero and this leads to a sparser model than those produced by a Gaussian prior.

Question 6

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &= \frac{1}{Z} p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) p(\mathbf{W}) \\ &= \frac{1}{Z} \mathcal{N}(\mathbf{W}|\mathbf{X}, \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{W}_0, \tau^2 \mathbf{I}) \end{aligned}$$

$\frac{1}{Z}$ is the normalising constant to ensure that the posterior is a probability density function by making the area under the graph equal to 1 this constant is called the evidence. We will ignore it for now.

By the probability density function of the multivariate normal distribution, we have:

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto \frac{1}{\sqrt{(2\pi)^N \sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{W})^T(\mathbf{Y} - \mathbf{X}\mathbf{W})\right) \\ &\quad \cdot \frac{1}{\sqrt{(2\pi)^N \tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mathbf{W} - \mathbf{W}_0)^T(\mathbf{W} - \mathbf{W}_0)\right) \end{aligned}$$

We ignore the normalising constants as we re-normalise with Z and then combine and multiply out the exponents.

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{W})^T(\mathbf{Y} - \mathbf{X}\mathbf{W}) - \frac{1}{2\tau^2}(\mathbf{W} - \mathbf{W}_0)^T(\mathbf{W} - \mathbf{W}_0)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \mathbf{X}\mathbf{W} + \mathbf{W}^T \mathbf{X}^T \mathbf{X}\mathbf{W}) \right. \\ &\quad \left. - \frac{1}{2\tau^2}(\mathbf{W}_0^T \mathbf{W}_0 - 2\mathbf{W}^T \mathbf{W}_0 + \mathbf{W}^T \mathbf{W})\right) \end{aligned}$$

We know the posterior will be Gaussian as both the likelihood and prior are so can assume it will take the form: $\exp((\mathbf{W} - \mu)^T \Sigma^{-1}(\mathbf{W} - \mu))$. If we multiply the exponent out we get a quadratic so we try and make the posterior we have look like this quadratic.

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto \exp\left(-\frac{1}{2\sigma^2} \mathbf{W}^T \mathbf{X}^T \mathbf{X} \mathbf{W} - \frac{1}{2\tau^2} \mathbf{W}^T \mathbf{W} \right. && \text{quadratic term} \\ &\quad \left. + \frac{1}{\sigma^2} \mathbf{W}^T \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{W}^T \mathbf{W}_0 \right. && \text{mixed term} \\ &\quad \left. - \frac{1}{2\sigma^2} \mathbf{Y}^T \mathbf{Y} - \frac{1}{2\tau^2} \mathbf{W}_0^T \mathbf{W}_0\right) && \text{constant term} \end{aligned}$$

By re-arranging and completing the square we can find both Σ^{-1} and $\Sigma^{-1}\mu$ in the quadratic and mixed terms respectively.

$$\begin{aligned} p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto \exp\left(-\frac{1}{2} \mathbf{W}^T \overbrace{\left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} + \frac{1}{\tau^2} \mathbf{I}\right)}^{\Sigma^{-1}} \mathbf{W} \right. && \text{quadratic term} \\ &\quad \left. + \mathbf{W}^T \underbrace{\left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{W}_0\right)}_{\Sigma^{-1}\mu} \right. && \text{mixed term} \\ &\quad \left. - \frac{1}{2\sigma^2} \mathbf{Y}^T \mathbf{Y} - \frac{1}{2\tau^2} \mathbf{W}_0^T \mathbf{W}_0\right) && \text{constant term} \end{aligned}$$

From these we can easily find Σ and μ .

$$\begin{aligned}\Sigma &= \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{I} \right)^{-1} \\ \Sigma^{-1} \mu &= \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{W}_0 \\ \mu &= \Sigma \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{W}_0 \right)\end{aligned}$$

Thus we have our posterior as a Gaussian with our values for Σ and μ :

$$\begin{aligned}p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &\propto \exp \left(-\frac{1}{2} (\mathbf{W} - \mu)^T \Sigma^{-1} (\mathbf{W} - \mu) \right) \\ &\propto \mathcal{N} \left[\left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{I} \right)^{-1} \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{W}_0 \right), \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{Y} + \frac{1}{\tau^2} \mathbf{I} \right)^{-1} \right]\end{aligned}$$

1.1.2 Non-parametric regression

Question 7 Parametric models assume that the distribution the data comes from is based on a finite, fixed set of parameters and models future predictions based off these parameters, they capture everything there is to know about the data. Non-parametric models do not make such assumptions about the model structure and instead infer structure from the data, they have parameters but these are not fixed in advance and there can be an infinite set of parameters.

Non-parametric models are more flexible and can represent a wider variety of data and will represent the data better if the assumptions made in the parametric model are incorrect but are less precise and accurate than parametric methods if the right assumptions are made.

Parametric models are often easier interpreted as they are simpler to transcribe and are also often faster to compute due to lacking the complexity and flexibility of the non-parametric models.

Question 8 As we use a Gaussian process we define this prior over functions and want our prior to put some constraints on the space of functions. The fact that its a Gaussian process means that for an arbitrary set of points x_i, \dots, x_j we assume that $p(f_i), \dots, f(x_j)$ is jointly Gaussian with mean 0 and covariance function k , from this we have the equation for our prior:

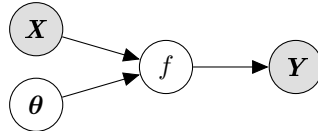
$$p(f|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(0, k(\mathbf{X}, \mathbf{X}))$$

The covariance function k allows us to set constraints from our assumptions about the mapping f . For example our assumption about the functions smoothness that if x_i and x_j are similar then we expect f_i and f_j to be similar too, with certain kernel functions for f we can only sample functions that have sufficient smoothness by ensuring that constraint.

Question 9 [TODO]

Question 10

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



[UNSURE]

- \mathbf{X} and $\boldsymbol{\theta}$ are independent
- f is conditionally dependent on both \mathbf{X} and $\boldsymbol{\theta}$
- \mathbf{Y} is conditionally dependent on f and conditionally independent of \mathbf{X} and $\boldsymbol{\theta}$

Question 11 [UNSURE] This marginalization shows the likelihood of the data we have observed over the function we are testing which is constrained by the prior.

There are two sources of uncertainty here, that associated with f in the prior and that associated with ϵ in the likelihood, these are independent and as such are merged by simply adding to form the covariance of the marginal likelihood Gaussian.

Leaving the θ on the left-hand side of the expression implies that we still have specific hyperparameters rather than undefined ones. It remains throughout the integral.

1.2 Practical

1.2.1 Linear regression

Question 12 Figure 1 represents our prior assumption over \mathbf{W} . Our prior assumption states that the distribution over \mathbf{W} be a Gaussian with a mean of 0 and the identity matrix as its covariance. As discussed in question 3 the likelihood is a Gaussian, therefore choosing a Gaussian prior allows us to obtain a Gaussian posterior.

In Figures 2, 3 and 4 we can see how our assumptions about \mathbf{W} change as we observe more data. Making the distribution over \mathbf{W} a posterior or our updated belief. These are paired with samples from our posterior that allow us to plot some sample functions. It is observable that the more data we observe, the more the uncertainty in our posterior reduces giving us a posterior that approaches the values of the actual parameters \mathbf{W} displayed as a white cross, and increasingly accurate sample functions.

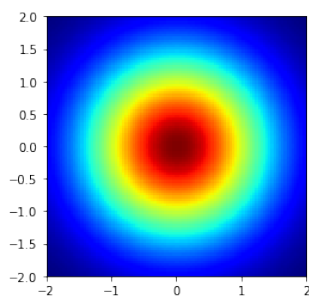


Figure 1: Prior for linear regression

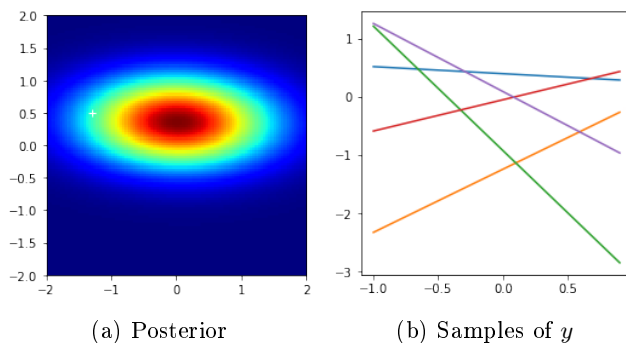


Figure 2: Linear regression - 1 observation

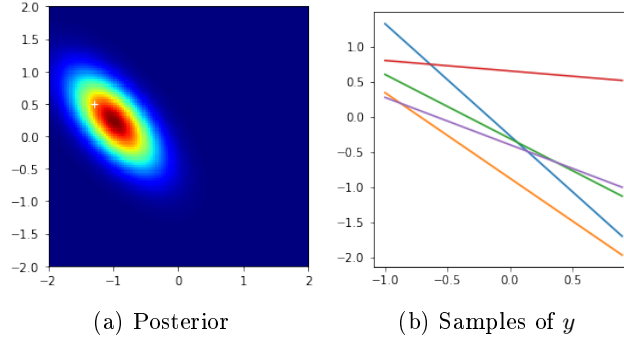


Figure 3: Linear regression - 2 observations

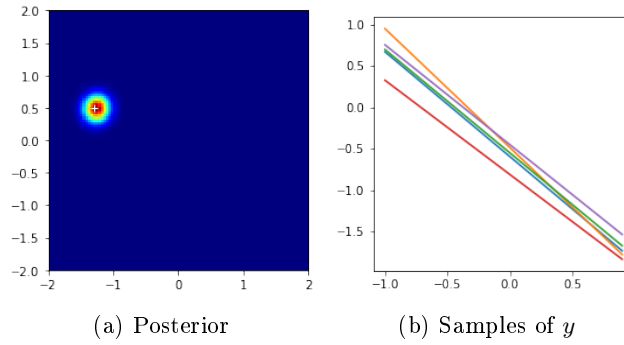


Figure 4: Linear regression - 15 observations

1.2.2 Non-parametric regression

Question 13 The covariance function, or kernel, encodes all the assumptions about the form of function we are modelling, and often represents some form of “distance” or similarity between the data.

Figure 5 shown represent samples from our GP-Prior created with a squared exponential covariance function. The length scale allows us to put constraints on how smooth the functions are. Small lengthscale values characterize functions that change quickly, whereas larger values characterize functions that change slowly. The three figures represent our GP-prior with different lengthscale values. We can observe that the high lengthscale value in Figure 5c gives us smooth functions while the lower lengthscale value in 5a gives us a functions with rapidly changing values.

The lengthscale therefore encodes our assumption of how smooth we think the underlying function will be.

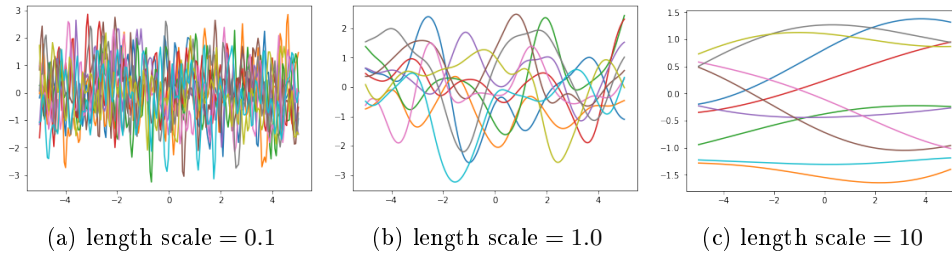


Figure 5: GP-prior samples

Question 14 Figure 6 shows the predictive posterior samples, or our updated prior after having observed data. While our prior showed functions that corresponded to our general prior assumptions, the posterior shows only the set of constrained functions that pass by the observed data points.

As the figure shows, this behavior is not desirable as it can lead to overfitting. Considering that there can be noise in the data, such as in Figure 6, overfitting will not enable us to find the underlying function, in this case a sine wave. We can observe that our uncertainty drops as we observe data but rises again as we get away from the observed data. This is due to our squared exponential covariance matrix that encodes an assumption in the form of the functions such that if two points x_1 and x_2 are close together, we expect their corresponding y values to be close too. Therefore the further away a new x value is from the x values of observed data points, the more uncertain we are about its y value.

Adding a diagonal covariance matrix to the covariance function encodes uncertainty in our model. As we can see in Figure 7b, there is a certain degree of error encoded in each data point observed, meaning that the underlying function does not pass through every observed data point, allowing us to obtain a smoother line that resembles far more the underlying sine function.

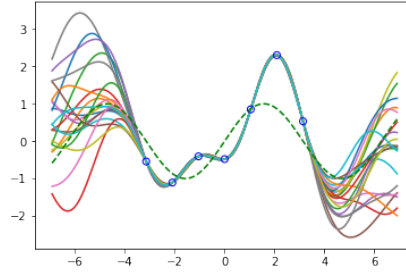


Figure 6: Predictive posterior samples

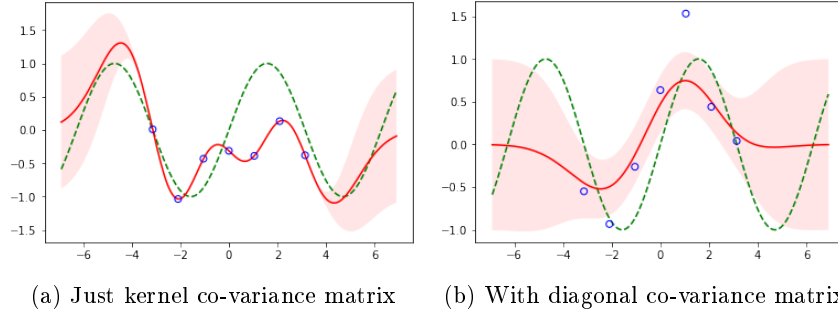


Figure 7: Predictive mean and variances

2 The posterior

2.1 Theory

Question 15 Specifying a prior assumption over \mathbf{X} allows us to encode our preference about the nature of the properties that \mathbf{X} should have. This assumption also constrains the variable \mathbf{W} as \mathbf{W} and \mathbf{X} have a simple relationship.

Question 16 We have encoded the assumption that all the dimensions in each variable of \mathbf{X} are independent and identically distributed as we have used a Gaussian with an identity matrix as its covariance matrix.

Question 17 [TODO]

2.1.1 Learning

Question 18 [TODO]

2.1.2 Practical optimisation

Question 19 1. We have from Question 17 that:

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

then the objective function is derived as follows using log rules:

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

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

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

2. To find the gradient of \mathcal{L} we will look at each term in turn:

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

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

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

We also have the rules that $\partial(\mathbf{X} \mathbf{Y}) = (\partial \mathbf{X}) \mathbf{Y} + \mathbf{X} (\partial \mathbf{Y})$ 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:

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

Using the earlier rules along with the rule that $\partial(\text{Tr}(\mathbf{X})) = \text{Tr}(\partial\mathbf{X})$ we can find the derivative of A :

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

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

$$\begin{aligned}\frac{\partial A}{\partial \mathbf{W}} &= \text{Tr} \left(\mathbf{Y} \mathbf{Y}^T \partial \left((\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I})^{-1} \right) \right) \\ \frac{\partial B}{\partial \mathbf{W}_{ij}} &= \text{Tr} \left(\mathbf{Y} \mathbf{Y}^T \left(-(\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{J}^{ij} \mathbf{W}^T + \mathbf{W} \mathbf{J}^{ijT}) (\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I}) \right) \right)\end{aligned}$$

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

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ij}} &= \frac{N}{2} \text{Tr} \left(\mathbf{Y} \left(-(\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{J}^{ij} \mathbf{W}^T + \mathbf{W} \mathbf{J}^{ijT}) (\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I}) \right) \mathbf{Y}^T \right) \\ &\quad + \frac{N}{2} \text{Tr} \left((\mathbf{W} \mathbf{W}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{J}^{ij} \mathbf{W}^T + \mathbf{W} \mathbf{J}^{ijT}) \right)\end{aligned}$$

2.1.3 Non-parametric

Question 20 [TODO]

2.2 Practical

2.2.1 Linear representation learning

Question 21 We minimised the objective function to get our maximum likelihood estimate for \mathbf{W}

$$\mathbf{W}' = \arg \min_{\mathbf{W}} \mathcal{L}(\mathbf{W})$$

To find the representation for \mathbf{X}' we must know the reverse of the linear mapping.

$$\mathbf{Y} = \mathbf{X}' \mathbf{W}'^T$$

$$\mathbf{X}' = \mathbf{Y} \mathbf{W}' (\mathbf{W}'^T \mathbf{W}')^{-1}$$

And then use minimization of our objective function using the non-linear conjugate gradient method to find \mathbf{X}' .

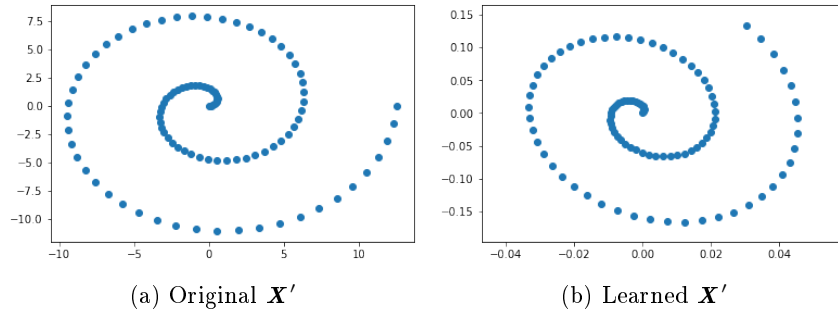


Figure 8: Representation Learning of \mathbf{X}'

Our learned representation of \mathbf{X}' is very close to the correct shape only rotated slightly, this is because the marginal likelihood we are maximising is invariant to any matrix transformation whose inverse is its own transposition i.e. any matrix for which $\mathbf{R}\mathbf{R}^T = \mathbf{I}$ (rotations

along with a few other transformations fall into this category). Therefore any of these transformations can be applied to \mathbf{W} without the marginal likelihood changing. We can see this by setting \mathbf{W} in the likelihood to $\mathbf{W}'\mathbf{R}$:

$$\begin{aligned}
\mathcal{L}(\mathbf{W}) &= - \sum_{i=1}^N \log (\mathcal{N}(y_i|0, (\mathbf{W}'\mathbf{R})(\mathbf{W}'\mathbf{R})^T + \sigma^2\mathbf{I})) \\
&= - \sum_{i=1}^N \log (\mathcal{N}(y_i|0, \mathbf{W}'\mathbf{R}\mathbf{R}^T\mathbf{W}'^T + \sigma^2\mathbf{I})) \\
&= - \sum_{i=1}^N \log (\mathcal{N}(y_i|0, \mathbf{W}'\mathbf{W}'^T + \sigma^2\mathbf{I}))
\end{aligned}$$