Coursework 1

Tuan Nguyen[1], Gizat Makhanov[2], Jeffrey Man Hong Chu[3]

[1] *97338*
[2] *97547*
[2] *16841*

November 2, 2018

In this assignment, we aim to criticise different key aspects of data modelling in Machine Learning : explore the supervised scenario with a model of a specific relationship between two different varieties; look at the unsupervised learning for some useful hidden structures or patterns of the data and apply the same approach as in regression task and use a Gaussian process prior over the mapping in particular non-parametric representation learning.

1 The Prior

Question 1

1. Gaussian likelihood function expresses how likely the outputs of the function, given input values x_i, match the target value y_i. Ideally, we strive for a function with zero difference, i.e. outputs and target values exactly match each other.

2. The reason for choosing a spherical covariance matrix for the likelihood is in simplifying work. Calculation of inverse of a non-diagonal matrix requires more computation and algebraic operations than inverse of a diagonal or spherical covariance matrix. With spherical covariance matrix we simply replace each element in the diagonal with its reciprocal.

Question 2

If each output is dependent on other data points, then we use the chain rule to express the joint distribution for a sequence of observations. The likelihood will then be:

$$p(\mathbf{Y}|f,\mathbf{X}) = \prod_{i=1}^{N} p(y_i|f,\mathbf{X}, \bigcap_{j=i+1}^{N} y_j) \tag{1}$$

Question 3

$$p(\mathbf{Y}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}x_n, \sigma^2\mathbf{I}) \tag{2}$$

Question 4

Given a likelihood, we can choose a prior in a particular distribution form, so that the posterior will also be in the same probability distribution family as the prior. The prior in this case is called a conjugate prior. Conjugate priors help reduce computation of the posterior to simple algebraic operations, generate the posterior exactly, and use an updated prior when calculating the following observations (sequential estimation deployment) by inputting posterior into prior.

Question 5

In the Gaussian distribution context, we can measure dissimilarity between one observation x and a set of observations x with mean μ and covariance σ^2, using the Mahalanobis distance. If covariance matrix is diagonal, then the distance between two points can be found using:

$$\sum_{i=1}^{N} \frac{x_i^2}{\sigma^2} \tag{3}$$

The intuition behind this is if a point x is closer to the center of mass of the spread out set, i.e. mean μ, then more probable that x belongs to x. Vice versa. We calculate that probability using a distance.

Question 6

The posterior probability according to Bayes' Theorem is defined as:

$$Posterior \propto Likelihood \times Prior \tag{4}$$

We can use this theorem to obtain parameters to build linear regression models for given input output data and predict future outcomes from similar observations. The posterior probability over parameters $\mathbf{w}$, given observations $\mathbf{x}$ and $\mathbf{y}$, would then be defined as:

$$p(\mathbf{w}|\mathbf{y},\mathbf{x}) = p(\mathbf{y}|\mathbf{w},\mathbf{x}) * p(\mathbf{w}) \tag{5}$$

Following Q4, we choose the likelihood and the prior to be Gaussians, to have a Gaussian posterior. The likelihood would then be a Gaussian of the form:

$$p(\mathbf{y}|\mathbf{w},\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{w}\mathbf{x},\sigma^2\mathbf{I}) \tag{6}$$

There's no prior knowledge about the prior, so we assume it's a Gaussian with zero mean:

$$p(\mathbf{w}) = \mathcal{N}(0,\Sigma) \tag{7}$$

As both the likelihood and the prior are Gaussians, we need to look at the general distribution form:

$$\begin{aligned}\mathcal{N}(x|\mu,\Sigma) &\propto e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)} \\ &= e^{-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x}e^{x^{\mathrm{T}}\Sigma^{-1}\mu}e^{-\frac{1}{2}\mu^{\mathrm{T}}\Sigma^{-1}\mu}\end{aligned} \tag{8}$$

If we write the exponents of the likelihood and the prior using the above form, our posterior equation transforms into:

$$\begin{aligned}&\underbrace{-\frac{1}{2\sigma^2}\mathbf{y}^{\mathrm{T}}\mathbf{y}\,\mathrm{A}}+\underbrace{\frac{1}{\sigma^2}\mathbf{y}^{\mathrm{T}}(\mathbf{x}\mathbf{w})\,\mathrm{B}} \\ &\qquad\underbrace{-\frac{1}{2\sigma^2}(\mathbf{x}\mathbf{w})^{\mathrm{T}}(\mathbf{x}\mathbf{w})-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\Sigma^{-1}\mathbf{w}}_{\mathrm{C}}\end{aligned} \tag{9}$$

Term A is constant, so we can look into terms B and C, which are linear and quadratic, respectively. By performing arithmetic operations on term C, we identify the covariance matrix of the posterior:

$$S^{-1} = \frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{x}+\Sigma^{-1} \tag{10}$$

We can then compare term B with the linear term of the density function of multivariate normal distributions, and derive mean μ.

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}S^{-1}\mu &= \frac{1}{\sigma^2}\mathbf{w}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{y} \\ \Rightarrow \mu &= \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{x}+\Sigma{-}1)^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}\end{aligned} \tag{11}$$

So given $\mathbf{x}$ and $\mathbf{y}$, the posterior would take the following Gaussian form:

$$\begin{aligned}p(\mathbf{w}|\mathbf{y},\mathbf{x}) = \mathcal{N}(\mathbf{w}|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{x}+\Sigma{-}1)^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}, \\ (\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{x}+\Sigma^{-1})^{-1}).\end{aligned} \tag{12}$$

Thus we derived our parameters for the linear regression model using only input output observation data and our assumptions about what these parameters are.

Question 7

In the parametric model, we know how the model look like, we assume some finite set of parameter. We need to find the parameter (w) then we predict new observation by using only the parameter.
In the non parametric model, we do not know how the model look like, we assume an infinite dimensional parameter. Usually we think of parameter as a function. We predict new observation by using the relationship betweeen data in training set and new observation.
Interpretability is the degree to which a human can consistently predict the model's result. The higher the interpretability of a model, the easier it is for us to comprehend why certain prediction were made. For parametric model, it is more interpretable, given that it contains sufficient number of data points with a clear and understandable formula. With the appropriate assumption, the predicted relationship will be less bias. In other way, non parametric model is less interpretable as in result of less data is required as compared with parametric model and the prediction cannot be made base on the formular.

Question 8

The prior is the distribution of function f. Because we dont have prior knowledge about function f, the mean is symmetrically set to 0. The covariance expresses that the more two points are similar, the more strongly correlated between two corresponding values.

Question 9

No. The prior only encodes the subset functions. By looking at the prior formulation, we can see that the prior is Gaussian distribution with zero mean and the covariance matrix expresses the relashionship between our data. So that this prior heavily depends on the data. As a result, it only encodes only subset functions.

Question 10

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

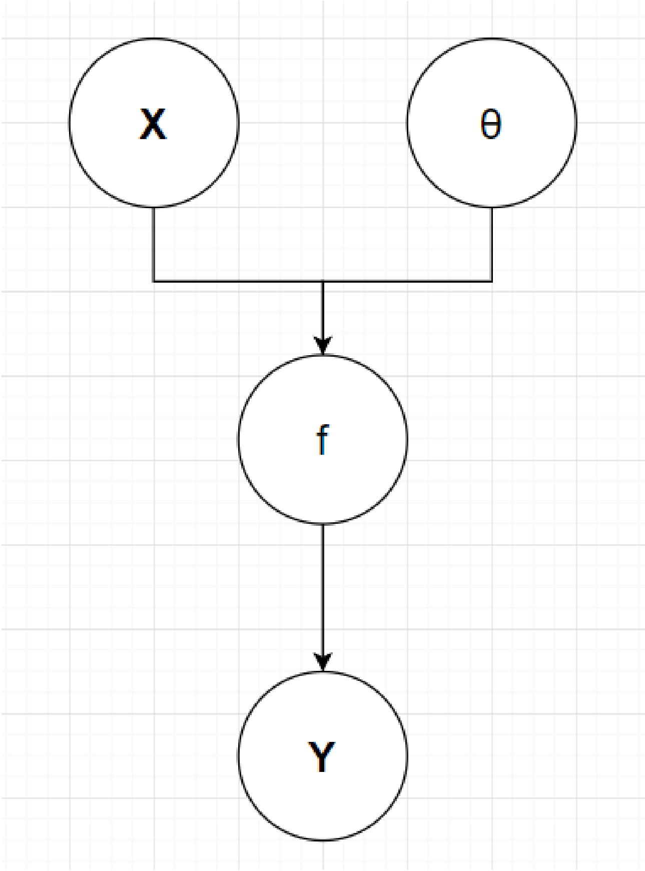

Figure 1: *Graphical model*

State about the model:

- $\mathbf{X}, \theta$, f is parameters of distribution of $\mathbf{Y}$
- $\mathbf{X}, \theta$ is parameters of distribution of f

Question 11

1. In the prior $p(f|\mathbf{X}, \theta)$, X is a parameter of f. And actually the Gram matrix in the prior expresses that the more two points are similar, the more strongly correlated between two corresponding values.
2. Because we marginalize the function f, which is something that we are uncertain, so the uncertainty actually filter through the marginalization.
3. θ is left on the left-hand side of the expression after marginalisation shows that θ is a parameter of distribution of $\mathbf{Y}$. It implies that the function f can encode all parameters θ.

Question 12

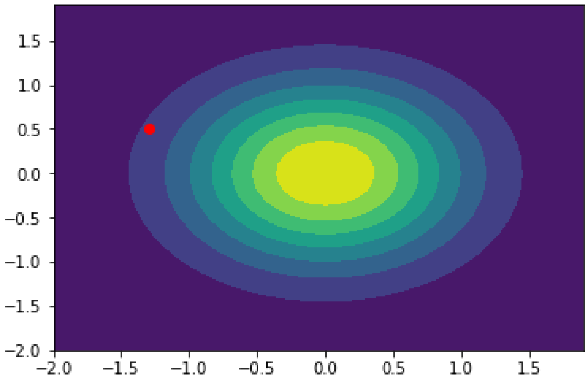

Figure 2: *Prior*

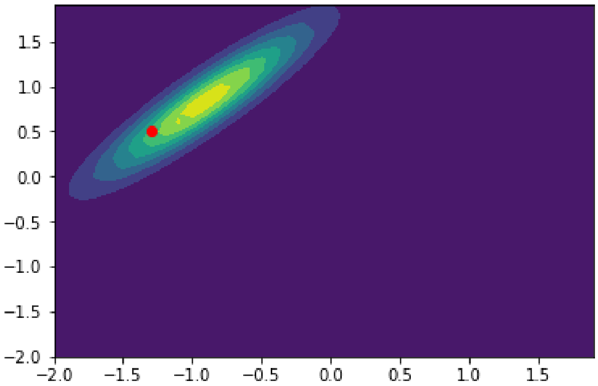

Figure 3: *Posterior add 1 point*

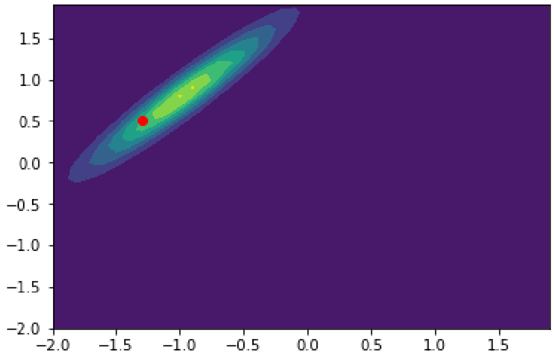

Figure 4: *Posterior add 2 points*

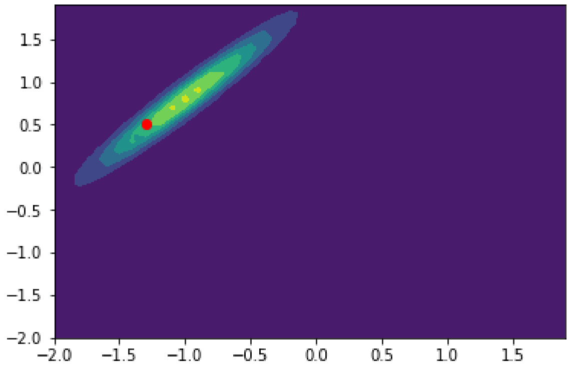

Figure 5: *Posterior add 3 points*

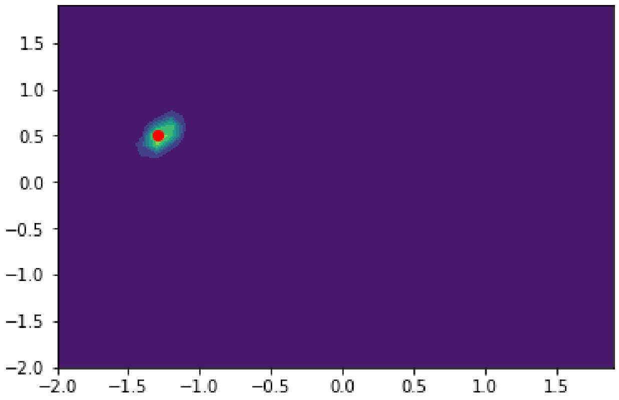

Figure 6: *Posterior add 20 points*

When we add more data points, the posterior becomes narrower because having more data points means that we are more certainly about our models. The more data point is added, the more the likelihood overwhelms the prior, so the posterior become narrower.

Question 13

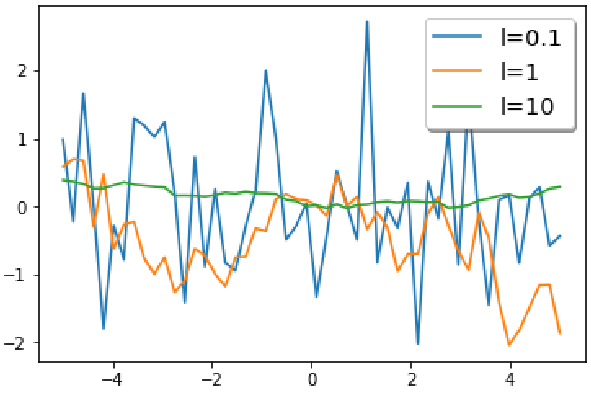

Figure 7: *Samples using different length-scale (0.1, 1, 10) for the squared exponential*

We can see that the prior becomes smoother when we increase length-scale. The length-scale also describes how smooth a function is. Small length-scale value means that function values can change quickly, large values characterize functions that change only slowly. Length-scale also determines how far we can reliably extrapolate from the training data.

Question 14

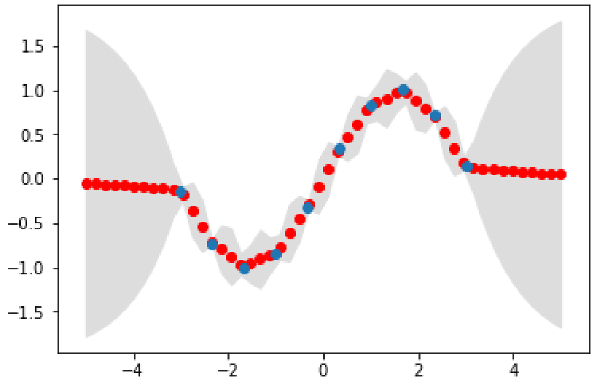

Figure 8: *Posterior*

When predicted point near the training set, the predicted value is more accurate and less uncertainty. In the posterior, the mean graph fits the training data with a bit uncertainty compared to random distribution in prior.
If we add a diagonal covariance matrix to the squared exponential, the uncertainty in each point is increased.

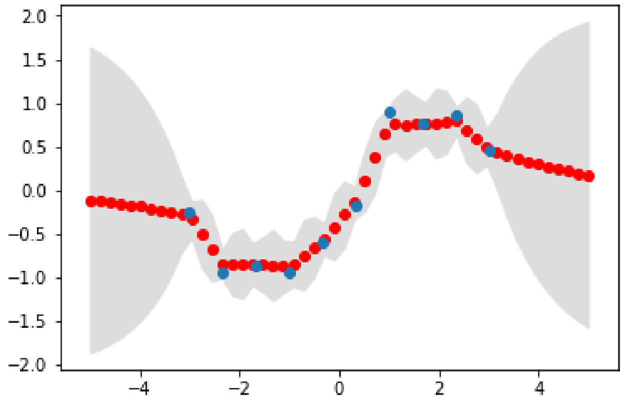

Figure 9: *Posterior with more covariance*

2 Posterior

Question 15

The belief helps us make assumption and the assumption expresses our preference

Question 16

The latent variable is what we are looking for and we have no prior knowledge about latent variable. Then, the latent variables are defined to be independent and Gaussian with unit variance for 2 main purposes:

- The prior is Gaussian then when we intergrate out latent variable, it is easier to calculate margianl distribution.
- The covariance matrix is identity, it is easier to do mathematic stuff such as invert matrix.

Question 17

$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})dX$

$= \int \mathcal{N}(\mathbf{Y}|\mathbf{WX},\sigma^2\mathbf{I})\mathcal{N}(\mathbf{X}|0,\mathbf{I})dX$

$= \int \frac{1}{(2\pi)^n\sigma^2}\exp\,(-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{WX})^\intercal(\mathbf{Y}-\mathbf{WX}) - \frac{1}{2}\mathbf{X}^\intercal\mathbf{X})dX$

Set A $= -\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{WX})^\intercal(\mathbf{Y}-\mathbf{WX}) - \frac{1}{2}\mathbf{X}^\intercal\mathbf{X}$

$=-\frac{1}{2\sigma^2}\mathbf{Y}^\intercal\mathbf{Y} + \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{WX} - \frac{1}{2}\mathbf{X}^\intercal(\frac{1}{\sigma^2}\mathbf{W}^\intercal\mathbf{W}+\mathbf{I})\mathbf{X}$

We have $-\frac{1}{2}(x-\mu_x)C_x^{-1}(x-\mu_x) = -\frac{1}{2}x^\intercal C_x^{-1}x + \mu_x^\intercal C_x^{-1}x - \frac{1}{2}\mu_x^\intercal C_z^{-1}\mu_x$

$\Rightarrow C_x^{-1} = \frac{1}{\sigma^2}\mathbf{W}^\intercal\mathbf{W}+\mathbf{I}$

We have $\mu_x^\intercal C_x^{-1} = \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{W} \Rightarrow \mu_x^\intercal = \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{W}C_x$

$\Rightarrow \mu_x^\intercal C_x^{-1}\mu_x = \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{W}(\mathbf{W}^\intercal\mathbf{W}+\sigma^2)^{-1}\mathbf{W}^\intercal\mathbf{Y}$

So that

$p(\mathbf{Y}|\mathbf{W}) \propto \exp A + \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{Y}))\int\mathcal{N}(\mathbf{X}|\mu_x,C_x)dX$

Where $A = (-\frac{1}{2}(-\frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{W}(\mathbf{W}^\intercal\mathbf{W}+\sigma^2)^{-1}\mathbf{W}^\intercal\mathbf{Y}$

We know that $\int\mathcal{N}(\mathbf{X}|\mu_x,C_x)dX = 1$ so that:

$p(\mathbf{Y}|\mathbf{W}) \propto$

$\exp\,(-\frac{1}{2}(-\frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{W}(\mathbf{W}^\intercal\mathbf{W}+\sigma^2)^{-1}\mathbf{W}^\intercal\mathbf{Y} + \frac{1}{\sigma^2}\mathbf{Y}^\intercal\mathbf{Y}))$

$=\exp\,(-\frac{1}{2}\mathbf{Y}^\intercal(\frac{1}{\sigma^2} - \frac{1}{\sigma^2}\mathbf{W}(\mathbf{W}^\intercal\mathbf{W}+\sigma^2)^{-1}\mathbf{W}^\intercal)\mathbf{Y})$

We use $(A+UCV)^{-1} = U(C^{-1}+VA^{-1}U)VA^{-1}$ with $A=\sigma^2 I, C=I, U=\mathbf{W}, V=\mathbf{W}^\intercal$

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

Question 18

1. Compare maximun likelihood type 1 (ML1), maximun likelihood type 2 (ML2), maximum a posterior (MAP)
   - ML1: Find the parameter by maximizing the likelihood function which expresses how the model fit the data.
   - MAP: Combination of prior (our belief about model before observing data) and ML1.
   - ML2: Marginalize 1 parameter and find another parameter by finding maximum of marginalized formulation.
2. When we add more data the maximum likelihood is sensitive to data which can cause overfitting, the MAP is kind of between the data and our belief (prior function). The more data is added, the more effect of prior is reduced.
3. Because the denominator of the middle equation, which is marginalized out w, does not depend on w.

Question 19

Objective function

$\mathcal{L}(\mathbf{W}) = \frac{ND}{2}\log 2\pi + \frac{N}{2}\log|C| + \frac{1}{2}\sum_{n=1}^{N} y_n^\intercal C^{-1}y_n$

The gradients of the objective with respect to the parameters $\frac{d\mathcal{L}}{d\mathbf{W}}$

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

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

We have

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

Where the matrix $\mathbf{J}_{ij}$ has all zero entries except for $(\mathbf{J}_{ij})_{ij} = 1$

We start by tackling the first term, the derivative of the log determinant

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

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

This is the first term, now we move to the second term. The derivative of the trace:

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

So the second term becoms,

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

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

Now we use the derivative of matrix inverse

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

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

Finally,

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

Question 20

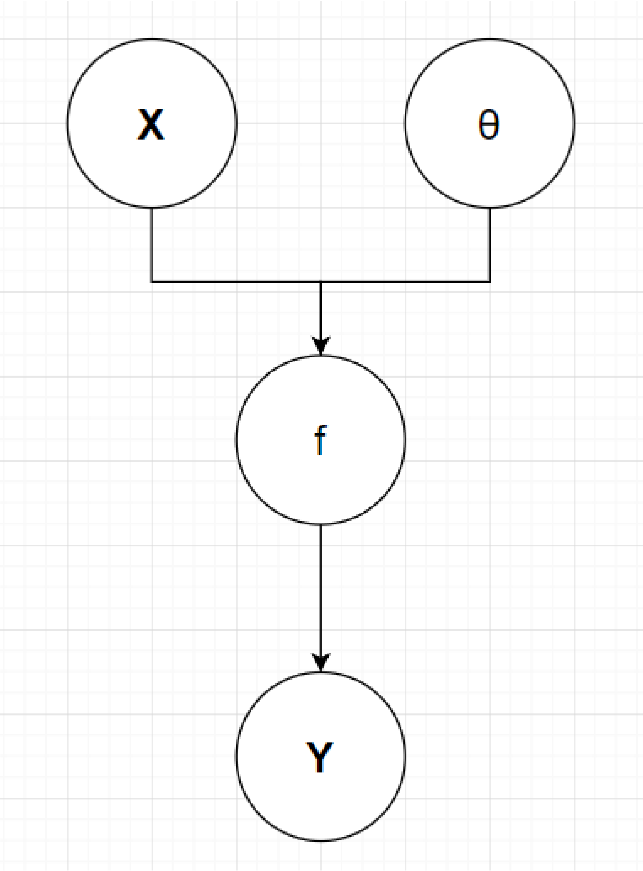

Figure 10: *Graphical model*

Looking at the model, we can see that $\mathbf{X} \to f \to \mathbf{Y}$. It means that **X** is the parameter of both distribution of f and **Y**, while f is the parameter of distribution of **Y** only so that margingalizing f is easier than marginalizing X.

Question 21

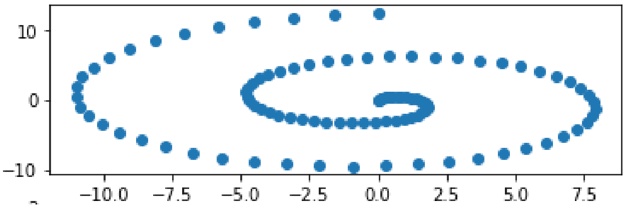

Figure 11: *Initial value to generate Y*

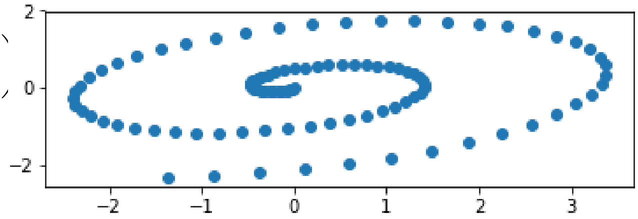

Figure 12: *Latent variable from Y*

We can see that the recovered data have the same shape with initial data, which means we successfully reduce the dimension of Y and get initial feature of X which is used to generate Y.

Question 22

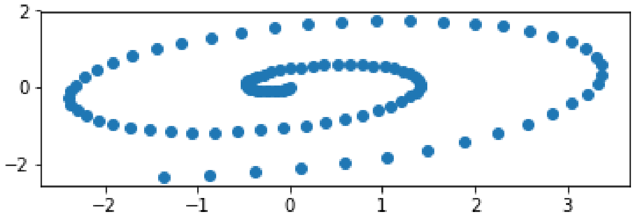

Figure 13: *Latent variable recovered from Y*

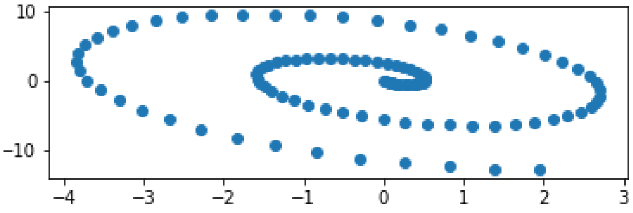

Figure 14: *Random two dimensional subspace*

The data in Figure 12 seems to be rotated and scaled up compared to Figure 11. The reason for that is because mapping function is linear so that when we map from 2 dimension to 10 dimension then we map back to 2 dimension, the data in 2 dimension can be rotated and scaled up compared to initial data. However the difference in latent data is that the covariance of projected data from Y to X is maximized.

Non-parametric presentation learning

Derivative objective function

$$\mathcal{L} = \frac{1}{2}\log|\mathbf{K}| + \frac{1}{2}tr(\mathbf{Y}^\intercal\mathbf{K}^{-1}\mathbf{Y})$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{X}_{ij}} = \frac{\partial\mathcal{L}}{\partial\mathbf{K}_{ij}}\frac{\partial\mathbf{K}_{ij}}{\partial\mathbf{X}_{ij}}$$
$$\frac{\partial\mathcal{L}}{\partial\mathbf{K}_{ij}} = \frac{1}{2}\frac{\partial\log|\mathbf{K}|}{\partial\mathbf{K}_{ij}} + \frac{1}{2}\frac{\partial tr(\mathbf{Y}^\intercal\mathbf{K}^{-1}\mathbf{Y})}{\partial\mathbf{K}_{ij}}$$
$$= \frac{1}{2}\frac{tr(\mathbf{K}^{-1}\partial\mathbf{K})}{\partial\mathbf{K}_{ij}} + \frac{1}{2}tr(\frac{\mathbf{Y}^\intercal\mathbf{K}^{-1}\mathbf{Y}}{\partial\mathbf{K}^{-1}}\frac{\partial\mathbf{K}^{-1}}{\partial\mathbf{K}_{ij}})$$
$$= \frac{1}{2}tr(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\mathbf{K}_{ij}}) + \frac{1}{2}tr(\mathbf{Y}^\intercal\mathbf{Y}\frac{-\mathbf{K}^{-1}\partial\mathbf{K}\mathbf{K}^{-1}}{\partial\mathbf{K}_{ij}})$$
$$= \frac{1}{2}tr(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\mathbf{K}_{ij}}) - \frac{1}{2}tr(\mathbf{Y}^\intercal\mathbf{Y}\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\mathbf{K}_{ij}}\mathbf{K}^{-1})$$

We have: $\frac{\partial\mathbf{K}}{\mathbf{K}_{ij}} = \mathbf{I}_{ij}$
The matrix $\mathbf{I}_{ij}$ has all zero entries except for $(\mathbf{I}_{ij})_{ij} = 1$

And $\frac{\mathbf{K}_{ij}}{\partial\mathbf{X}_{ij}} = \frac{\sigma^2\exp-\frac{(x_i-x_j)^\intercal(x_i-x_j)}{l^2}}{\partial\mathbf{X}_{ij}} = \frac{-2\mathbf{K}_{ij}(x_i)_j}{l^2}$
where $(x_i)_j$ is j^{th} dimension of x_i

$$\Rightarrow \frac{\partial\mathbf{L}}{\partial\mathbf{X}_{ij}} = \frac{\mathbf{K}_{ij}(x_i)_j}{l^2}(tr(\mathbf{Y}^\intercal\mathbf{Y}\mathbf{K}^{-1}\mathbf{I}_{ij}\mathbf{K}^{-1}) - tr(\mathbf{K}^{-1}\mathbf{I}_{ij}))$$

Similarly, we have derivative of hyperparameters

$$\frac{\partial\mathcal{L}}{\partial\sigma} = \frac{\partial\mathcal{L}}{\partial\mathbf{K}_{ij}}\frac{\partial\mathbf{K}_{ij}}{\partial\sigma}$$
$$\frac{\partial\mathbf{K}_{ij}}{\partial\sigma} = \frac{\partial(\sigma^2\exp-\frac{(x_i-x_j)^\intercal(x_i-x_j)}{l^2})}{\partial\sigma} = \frac{2\mathbf{K}_{ij}}{\sigma}$$
$$\Rightarrow \frac{\partial\mathbf{L}}{\partial\sigma} = \frac{\mathbf{K}_{ij}}{\sigma}(tr(\mathbf{K}^{-1}\mathbf{I}_{ij}) - tr(\mathbf{Y}^\intercal\mathbf{Y}\mathbf{K}^{-1}\mathbf{I}_{ij}\mathbf{K}^{-1}))$$

And

$$\frac{\partial\mathcal{L}}{\partial l} = \frac{\partial\mathcal{L}}{\partial\mathbf{K}_{ij}}\frac{\partial\mathbf{K}_{ij}}{\partial l}$$
$$\frac{\partial\mathbf{K}_{ij}}{\partial l} = \frac{\partial(\sigma^2\exp-\frac{(x_i-x_j)^\intercal(x_i-x_j))}{l^2}}{\partial l} = \frac{2\mathbf{K}_{ij}x_i^\intercal x_j}{l^3}$$
$$\Rightarrow \frac{\partial\mathbf{L}}{\partial l} = \frac{2\mathbf{K}_{ij}x_i^\intercal x_j}{l^3}(tr(\mathbf{K}^{-1}\mathbf{I}_{ij}) - tr(\mathbf{Y}^\intercal\mathbf{Y}\mathbf{K}^{-1}\mathbf{I}_{ij}\mathbf{K}^{-1}))$$

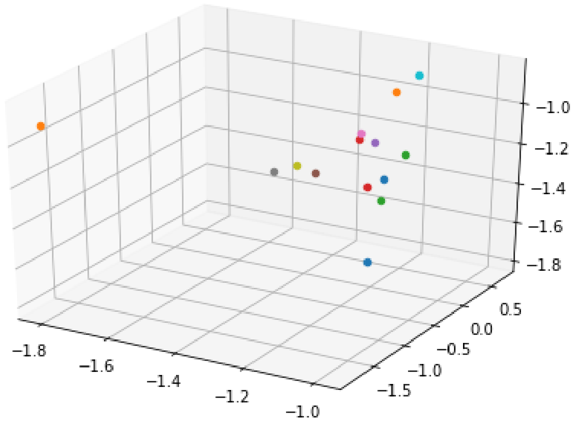

Figure 15: *1 frame of motion capture data of a person running in 3D*

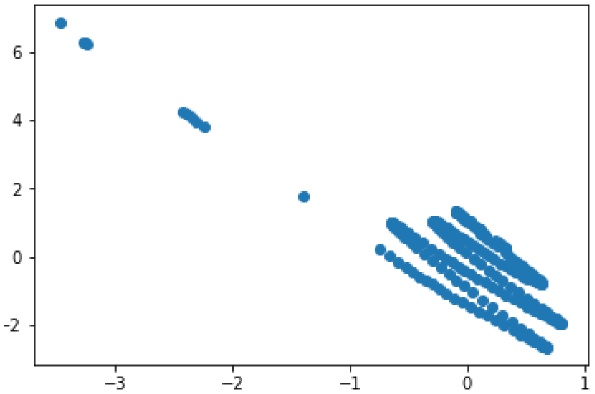

Figure 16: *Latent variable restored from the data*

Looking at the Figure 16, we can see the zigzag in the restored structure because the running motion seems to be repeated so we can see the similarity between points in the latent structure of running motion training data.
In order to find latent variable, we use maximum likelihood type 2. Instead of marginalizing out the latent in the linear latent variable model, in the non-linear latent model, we marginalize out the parameter, then we replace the linear kernel of X by non-linear kernel such as squared exponential covariance function. Then we find the latent variable by maximizing the distribution of data.
In case of linear mapping, the shape can be rotated or scaled up, and that is it. However in non-linear case, it can be much powerful than that, the non-linear transformation. For example, in our case, we can restore the two-dimensional latent variable from motion capture data.

3 Evidence

Question 23

1. This is the simplest possible model. Because this model has no parameter so we do not need to find the parameter. As a result, it is simple model
2. This is the most complex model. Because our data domain has 512 different elements and the model places probability mass over the whole data equally, so the probability mass in each data point is so small. As a result, this is complex model.

Question 24

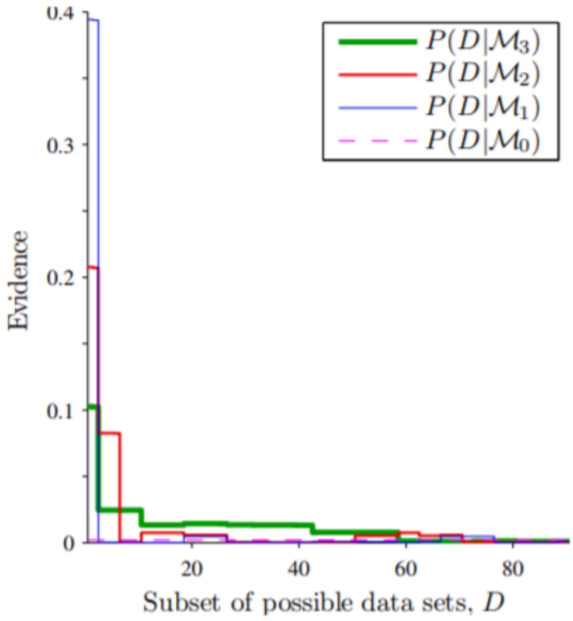

Figure 17: *Plot of evidence for data sets for the models [1]*

Looking at the figure, we can see that distribution of model 0 focus on the whole data set, while the distribution of other models only focus on subset of data. Except for the model 0, the more parameter that the model has, the bigger region that distribution of model focus on. Because the distributions are normalized so the more the model spread out the data, the smaller the probability mass in each data point is or the more we are uncertainty about the model at this point. Consequently, while the data cannot be fitted well by the simple model, the more complex model focuses its predictive probability on a large range of data sets and so each point only has relatively small probability.

4 Final thoughts

After doing the coursework, we know that there are two main types of machine learning including supervised learning and unsupervised learning. While in supervised learning, we find the relationship between two variates which are the data and label, in unsupervised learning, we find the hidden structure of the data. And we understand how to solve these two main problems. In addition, we know how to integrate out the belief with observation using a simple set of rule. Let's use Bayesian viewpoint to find our difficulty in machine learning course. Our prior before attending this course is that it is not hard because we got distinction in the bachelor degree or we attended machine learning "cookbook" course online. However, after the data is added, which is attending more and more lectures, our posterior now becomes that this subject is so hard. The proof for this is that we have to spend most of our time on machine learning subject only. And even when we are kind of understand the lecture, it is still hard to understand the questions in the coursework.

References

[1] Iain Murray and Zoubin Ghahramani. A note on the evidence and Bayesian Occam's razor Technical Report GCNU-TR 2005-003, August 2005.