1.

a) (This was in previous year’s part 1 tutorial)

i)

dimensions:

ii.

Dim: R^ExE

Dim: R^ExD

Dim: R^ExD

b)

i.

likelihood

prior

evidence/normalization

ii.

(Not examined anymore?)

Occams Razor: choose thing with least assumptions.

In model selection: Choose model with fewest hyperparameters, least complexity, as this requires less assumptions about the data.

iii.

Bayesian inference considers the parameters as a distribution (instead of point estimates) and we weight each model (defined by its parameters) by a probability.

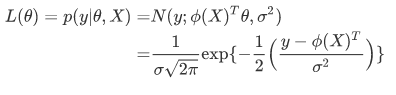
By making an initial simple guess at the model (prior), we can then compute a posterior distribution (a new guess).

The posterior may now imply a more complex model, but only since we have the data to back this up. So we’ve not made any assumptions to increase the complexity of the model.

So this follows Occam's razor: our model always uses the least assumptions. Complexity is only increased by evidence not by assumption.

c)

i.



\begin{align\*}

L(\theta) = p(y|θ,X)

=& N(y;ϕ(X)^{T}\theta,σ^{2})\\

=& \frac{1}{\sigma\sqrt{2\pi}} \exp \{ -\frac{1}{2} \Big(\frac{y - ϕ(X)^{T}}{\sigma^{2}} \Big)\} \\

\end{align\*}

ii.

Overfitting is having parameters that (may be very large/complex) and very closely model the training data. Problem is that by fitting too closely to training data, we assume that the data has little/no noise, but instead we may be fitting the noise and it can be inaccurate on unseen data

iii.

Instead of using the likelihood function directly (that’s MLE), we instead combine this with a prior distribution – an initial guess about what we suspect the parameters to be. We can then combine these together using bayes rule to infer our estimate about the parameters (so choose a model).

So MAP has this extra information about what we initially suspect the parameters to be. This prevents wild erratic guesses at the parameters, as it grounds them. Making very complex parameters very unlikely. So it is more robust to overfitting.

Alt answer:

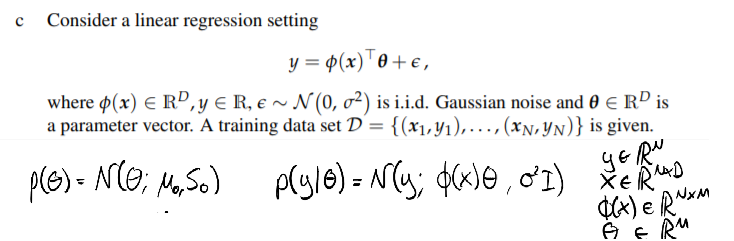
MAP is MLE with a regularization term, such that we can modify the objective function to add a penalty term which prevents its training loss from reaching 0 i.e. perfect fit to the data. This also delays the dramatic increase in test loss, at which point the model has begun to overfit to the training data. Both of these points help to avoid overfitting at a more clear cutting point than MLE.

iv.

Normal/Gaussian distribution.

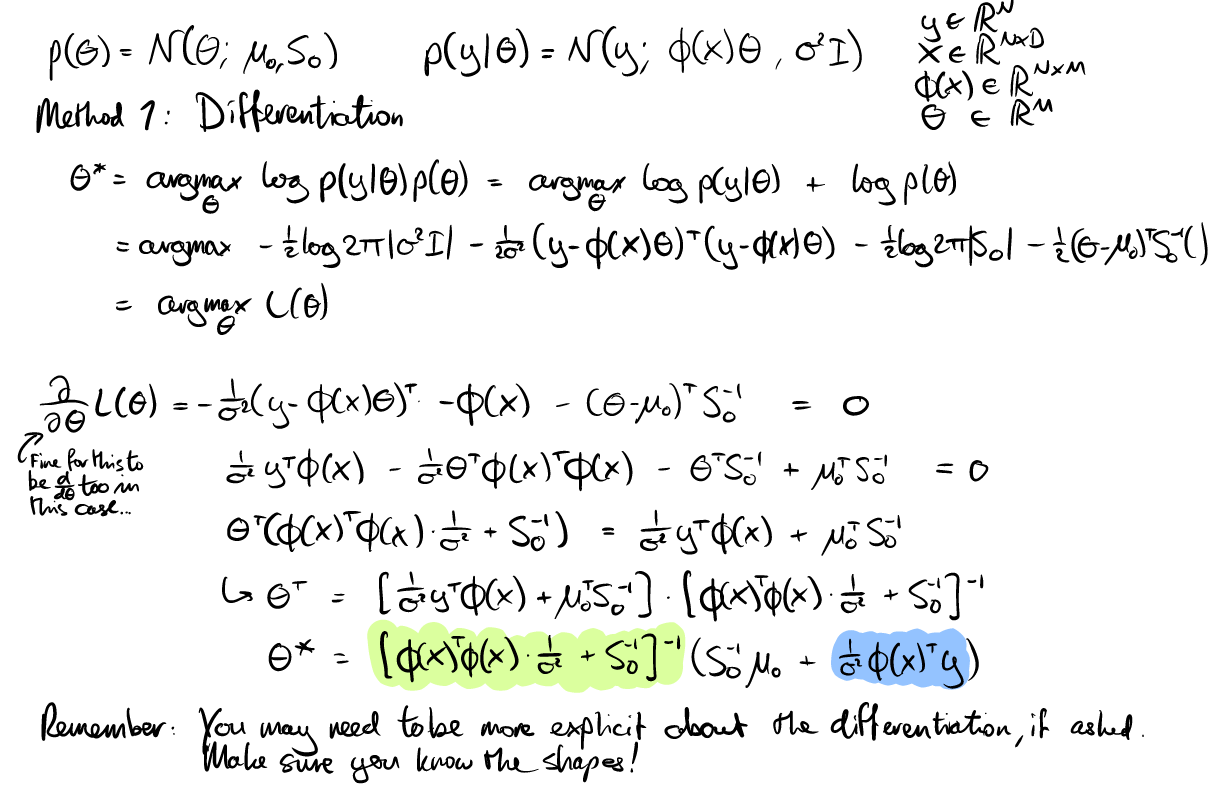
Since (as in i.) the likelihood is a Gaussian. One conjugate prior here therefore is a gaussian. Being a conjugate prior therefore implies the posterior will also be a gaussian, so it can be computed in closed form.

v.

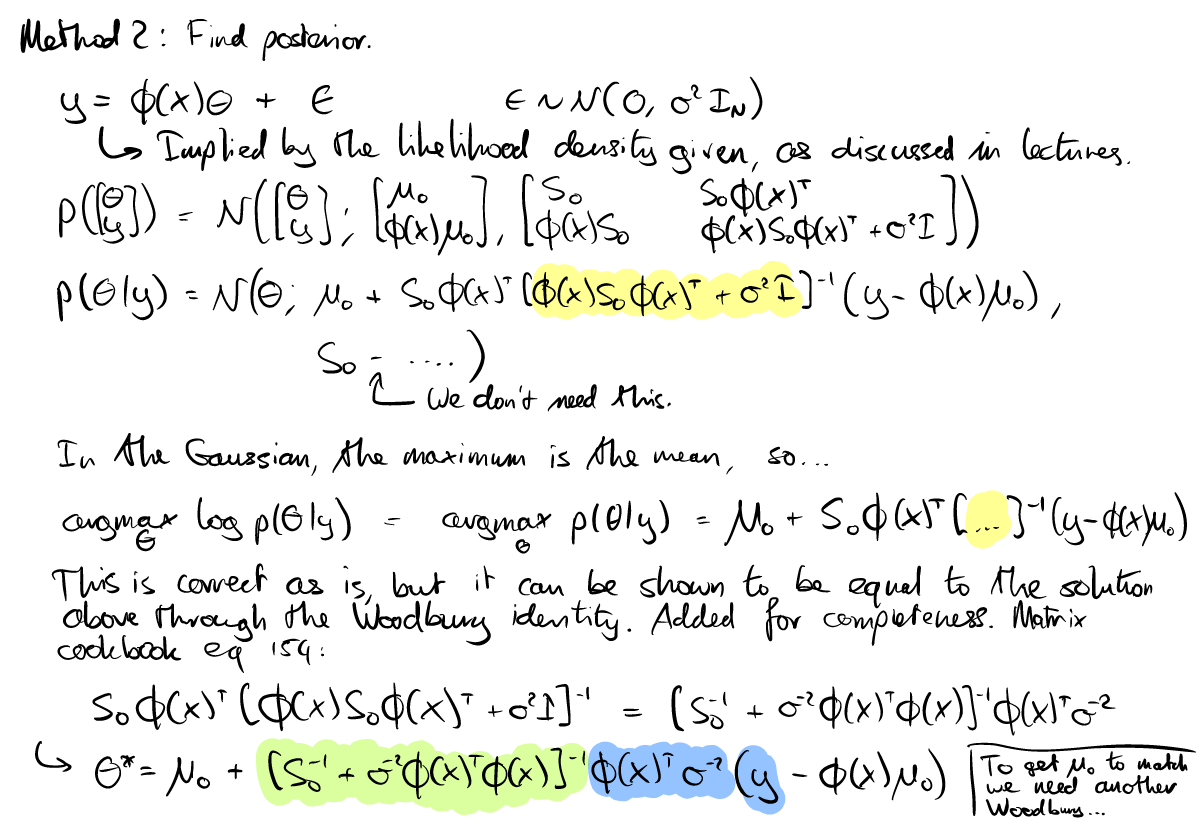


Mark just answered this question on Materials:

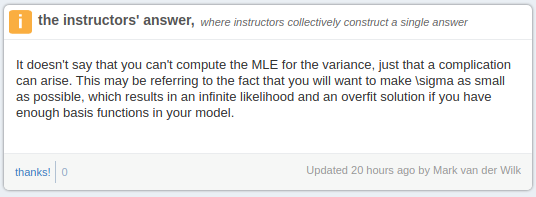
<https://materials.doc.ic.ac.uk/view/2021/70015/Course%20Material/64>



OR



vi.



2.

a)

i.

Set = 0 and you’ll get

ii.

* + - * Asymptotically consistent estimator
      * Gives data high probability????
      * Overfits the same as squared loss

From el google:

* + - * Unbiased
      * Efficient (we haven’t been taught this)
      * Asymptotically Consistent

iii.

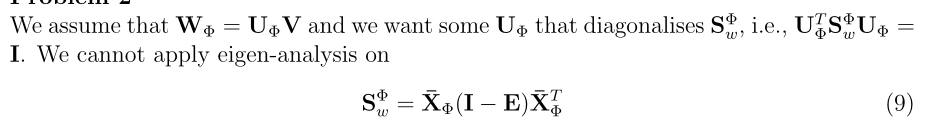
(Not examined anymore I think)

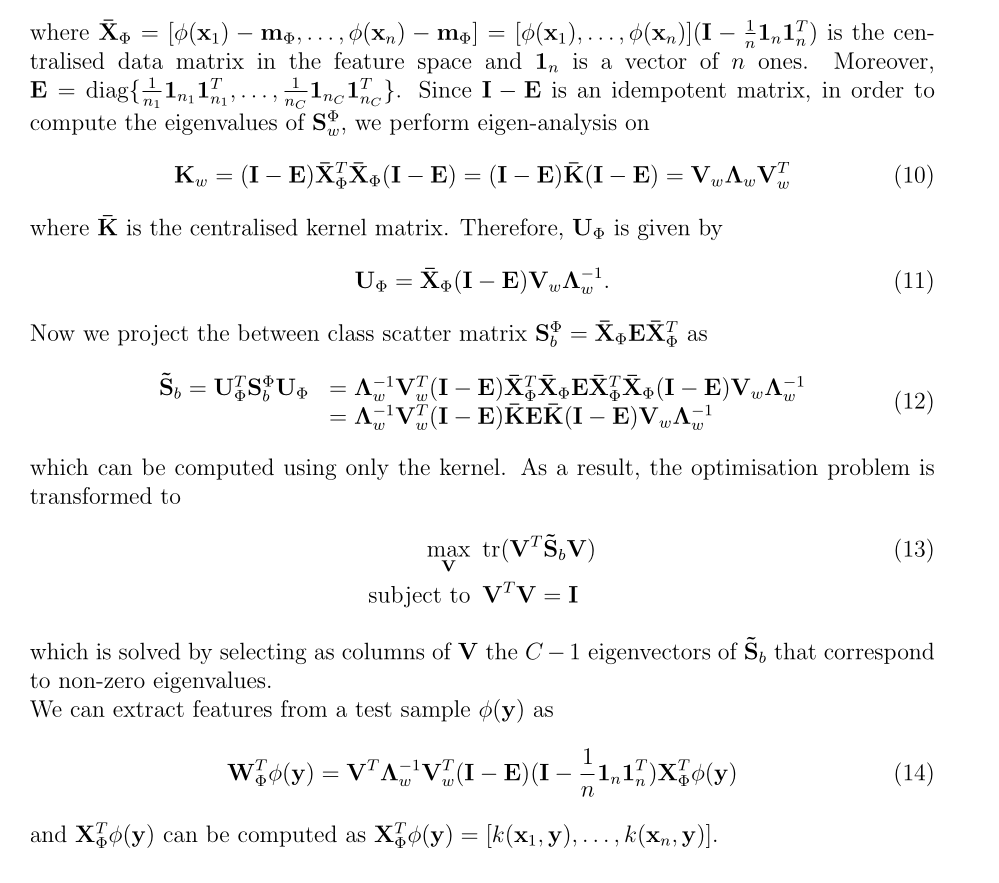
b) 6.5 in the MML book

c) nahh

3.

a) same as 19-20 3a





b) same as 19-20 3b

4.

Same as 19-20 2

a)

See 3.1.2 in lecture notes.

b)

Same thing but derivative of langrian is now:

$$\frac{\partial L}{ \partial w} = S\_{t} w - \sum\_{i} a\_{i} y\_{i} x\_{i} $$

so w is now: $$w = S\_{t}^{-1} \sum\_{i} a\_{i} y\_{i} x\_{i} $$

Then the dual problem becomes:

$$\sum \sum a\_{i} y\_{i} x\_{i}^{T} S\_{t}^{-1} x\_{j} y\_{j} a\_{j} $$

so we get the kernel as required

Idk about what to do if $$S\_{t}$$ is singular tho???