Homework 5 Solutions

Deadline: Wednesday, Nov. 14, at 11:59pm.

Submission: You need to submit two files:

1. Your solutions to Questions 1 and 2 as a PDF file, `hw5_writeup.pdf`, through MarkUs[1]. *(If you submit answers to Question 3, we will give feedback, but you will get the points for free; see below.)*
2. Your completed Python code for Question 1, as `q1.py`.

Neatness Point: One of the 10 points will be given for neatness. You will receive this point as long as we don't have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

Collaboration. Weekly homeworks are individual work. See the Course Information handout[2] for detailed policies.

1. **[3pts] Gaussian Discriminant Analysis.** For this question you will build classifiers to label images of handwritten digits. Each image is 8 by 8 pixels and is represented as a vector of dimension 64 by listing all the pixel values in raster scan order. The images are grayscale and the pixel values are between 0 and 1. The labels y are $0, 1, 2, \cdots, 9$ corresponding to which character was written in the image. There are 700 training cases and 400 test cases for each digit; they can be found in `a2digits.zip`.

   Starter code is provided to help you load the data (`data.py`). A skeleton (`q1.py`) is also provided for each question that you should use to structure your code.

   Using maximum likelihood, fit a set of 10 class-conditional Gaussians with a separate, full covariance matrix for each class. Remember that the conditional multivariate Gaussian probability density is given by,

$$p(\mathbf{x} \mid y = k, \boldsymbol{\mu}, \Sigma_k) = (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{1}$$

   You should take $p(y = k) = \frac{1}{10}$. You will compute parameters μ_{kj} and Σ_k for $k \in (0...9), j \in (1...64)$. You should implement the covariance computation yourself (i.e. without the aid of 'np.cov'). *Hint: To ensure numerical stability you may have to add a small multiple of the identity to each covariance matrix. For this assignment you should add* $0.01\mathbf{I}$ *to each matrix.*

   (a) **[1pt]** Using the parameters you fit on the training set and Bayes rule, compute the average conditional log-likelihood, i.e. $\frac{1}{N} \sum_{i=1}^{N} \log(p(y^{(i)} \mid \mathbf{x}^{(i)}, \theta))$ on both the train and test set and report it.

[1] `https://markus.teach.cs.toronto.edu/csc411-2018-09`

[2] `http://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/syllabus.pdf`

(b) **[1pt]** Select the most likely posterior class for each training and test data point as your prediction, and report your accuracy on the train and test set.

(c) **[1pt]** Compute the leading eigenvectors (largest eigenvalue) for each class covariance matrix (can use `np.linalg.eig`) and plot them side by side as 8 by 8 images.

Report your answers to the above questions, and submit your completed Python code for `q1.py`.

2. **[2pts] Categorial Distribution.** Let's consider fitting the categorical distribution, which is a discrete distribution over K outcomes, which we'll number 1 through K. The probability of each category is explicitly represented with parameter θ_k. For it to be a valid probability distribution, we clearly need $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. We'll represent each observation $\mathbf{x}$ as a 1-of-K encoding, i.e, a vector where one of the entries is 1 and the rest are 0. Under this model, the probability of an observation can be written in the following form:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}.$$

Denote the count for outcome k as N_k, and the total number of observations as N. In the previous assignment, you showed that the maximum likelihood estimate for the counts was:

$$\hat{\theta}_k = \frac{N_k}{N}.$$

Now let's derive the Bayesian parameter estimate.

(a) **[1pts]** For the prior, we'll use the Dirichlet distribution, which is defined over the set of probability vectors (i.e. vectors that are nonnegative and whose entries sum to 1). Its PDF is as follows:

$$p(\boldsymbol{\theta}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1}.$$

A useful fact is that if $\boldsymbol{\theta} \sim \text{Dirichlet}(a_1, \ldots, a_K)$, then

$$\mathbb{E}[\theta_k] = \frac{a_k}{\sum_{k'} a_{k'}}.$$

Determine the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$, where $\mathcal{D}$ is the set of observations. From that, determine the posterior predictive probability that the next outcome will be k.

Solution: First, we want to find an expression for the posterior distribution, $p(\boldsymbol{\theta} \mid \mathcal{D})$. The posterior distribution can be found using Bayes rule.

$$posterior = \frac{prior \cdot likelihood}{evidence}$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta}) \cdot p(\mathcal{D} \mid \boldsymbol{\theta})}{p(\mathcal{D})}$$

Since $p(\mathcal{D})$ is consistent for every value of $\boldsymbol{\theta}$, it can be disregarded as a normalising constant. Thus to find an expression *proportional to* the posterior distribution, we only need to find expressions for the prior, $p(\boldsymbol{\theta})$, and for the likelihood, $p(\mathcal{D} \mid \boldsymbol{\theta})$.

For a Dirichlet distribution we know that:

$$p(\boldsymbol{\theta}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1}.$$

So we only need to find an expression for the likelihood, $p(\mathcal{D} \mid \boldsymbol{\theta})$. The likelihood of all classes given the data can be written as:

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = p(x^{(1)}, ..., x^{(n)} \mid \boldsymbol{\theta})$$

In the question we are told that (D) is a set of observations. we denote the individual observations as $x^{(n)}$ where n is the observation number. We can apply the chain rule to reduce this statement.
Under the assumption that all of the data is independent and identically distributed, it is implied that any value $x^{(i)}$ will be independent of any other value $x^{(j)}$. The chained expression simplifies to:

$$p(x^{(N)}.....x^{(1)} \mid \boldsymbol{\theta}) = p(x^{(N)} \mid \boldsymbol{\theta}) \cdot p(x^{(N-1)} \mid \boldsymbol{\theta}) \cdots p(x^{(1)} \mid \boldsymbol{\theta})$$

This can be expressed in the following notation:

$$p(x^{(N)}.....x^{(1)} \mid \boldsymbol{\theta}) = \prod_{(n=1)}^{N} p(x^{(n)} \mid \boldsymbol{\theta})$$

The formula for a Dirichlet distribution for a 1-of-k encoding is given as:

$$p(x \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}$$

So for K classes, and N samples we can express the likelihood as

$$likelihood = p(\mathcal{D} \mid \boldsymbol{\theta}) = p(x^{(N)}.....x^{(1)} \mid \boldsymbol{\theta}) = \prod_{(n=1)}^{N} \left(\prod_{k=1}^{K} \theta_k^{x_k^{(n)}} \right)$$

Now we know that the posterior distribution is proportional to the product of the likelihood and the prior.

$$posterior \propto prior \cdot likelihood$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\boldsymbol{\theta}) \cdot p(\mathcal{D} \mid \boldsymbol{\theta})$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1} \cdot \prod_{(n=1)}^{N} \left(\prod_{k=1}^{K} \theta_k^{x_k^{(n)}} \right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \theta_k^{a_k - 1} \cdot \prod_{(n=1)}^{N} \left(\prod_{k=1}^{K} \theta_k^{x_k^{(n)}} \right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \left(\theta_k^{a_k - 1} \cdot \prod_{(n=1)}^{N} \theta_k^{x_k^{(n)}} \right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \left(\theta_k^{a_k - 1} \cdot \theta_k^{\sum_{n=1}^{N} x_k^{(n)}} \right)$$

We can combine the exponents like so:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \left(\theta_k^{a_k - 1 + \sum_{n=1}^{N} x_k^{(n)}} \right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \left(\theta_k^{a_k - 1 + N_k} \right)$$

Where we express $\sum_{n=1}^{N} x_k^{(n)}$ as N_k.

We can see that the posterior estimation is still Dirichlet distributed.

To determine the posterior predictive probability we must find the expectation of the posterior. We are given that if θ is approximately Dirichlet (i.e. $p(\boldsymbol{\theta}) \propto \theta_1^{a_1 - 1} \cdots \theta_K^{a_k - 1}$), resulting in an expectation of:

$$\mathbb{E}[\theta_k] = \frac{a_k}{\sum_{k'} a_{k'}}$$

Our posterior is also approximately Dirichlet (i.e. $p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \theta_1^{a_1 + N_1} \cdots \theta_K^{a_k + N_k}$). Our expectation of the posterior becomes:

$$\mathbb{E}[\theta_k \mid \mathcal{D}] = \frac{a_k + N_k}{\sum_{k'} N_{k'} + a_{k'}}$$

(b) **[1pt]** Still assuming the Dirichlet prior distribution, determine the MAP estimate of the parameter vector $\boldsymbol{\theta}$. For this question, you may assume each $a_k > 1$.

Solution: For the MAP estimate, we seek to find the optimum point of the posterior. We just showed that our posterior distribution is:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k'=1}^{K} \left(\theta_{k'}^{a_{k'} - 1 + N_{k'}} \right)$$

For convenience we can reframe this problem as maximising the log-likelihood of the posterior probability:

$$\arg\max_{\theta_{k'}} \left(\prod_{k'=1}^{K} \left(\theta_{k'}^{a_{k'} - 1 + N_{k'}} \right) \right) \equiv \arg\max_{\theta_{k'}} \log \left(\prod_{k'=1}^{K} \left(\theta_{k'}^{a_{k'} - 1 + N_{k'}} \right) \right)$$

Our objective function becomes $\sum_{k'=1}^{K} (a_{k'} - 1 + N_{k'}) \cdot \log \theta_{k'}$ And we know that $\sum_{k'=1}^{K} \theta_{k'} = 1$.

When we want to solve optimisation problems with constraints we can use Lagrange multipliers.

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda \cdot g(x, y)$$

Where f(x,y) is the function that we wish to optimise, and g(x,y) are the constraints we are subject to.

Since we want to find where the function is optimal, we set the derivative to zero.

$$f(x, y) = \lambda \cdot g(x, y)$$

For our case, our only parameter is θ_k.

$$f(\theta_k) = \lambda \cdot g(\theta_k)$$

Taking the derivative of $f(\theta_{k'})$ we arrive at the following:

$$\frac{\partial}{\partial \theta_k} \cdot f(\theta_k) = \frac{\partial}{\partial \theta_k} \cdot \sum_{k'=1}^{K} (a_{k'} - 1 + N_{k'}) \cdot \log \theta_{k'}$$

$$\frac{\partial}{\partial \theta_k} \cdot f(\theta_k) = \frac{\partial}{\partial \theta_k} \cdot (a_k - 1 + N_k) \cdot \log \theta_k$$

This is equivalent to:

$$\frac{\partial}{\partial \theta_k} \cdot f(\theta_k) = (a_k - 1 + N_k) \cdot \frac{\partial}{\partial \theta_k} (\log \theta_k)$$

$$\frac{\partial}{\partial \theta_k} \cdot f(\theta) = \frac{(a_k - 1 + N_k)}{\theta_k}$$

For our constraint we can solve as follows:

$$\frac{\partial}{\partial \theta_k} \cdot g(\theta_k) = \frac{\partial}{\partial \theta_k} \cdot \sum_{k=1}^{K} \theta_k = 1$$

Plugging both $f'(\theta_k)$ and $g'(\theta_k)$ into $f'(\theta_k) = \lambda \cdot g'(\theta_k)$ yields:

$$\frac{(a_k - 1 + N_k)}{\theta_k} = \lambda \cdot 1$$

$$\theta_k = \frac{(a_k - 1 + N_k)}{\lambda}$$

Recall that our function, $g(\theta_k) = \sum_{k=1}^{K} \theta_k = 1$. Using this knowledge we can do the following:

$$\sum_{k=1}^{K} \theta_k = \frac{\sum_{k=1}^{K} (a_k - 1 + N_k)}{\lambda} = 1$$

Therefore, by rearranging we find that:

$$\lambda = \sum_{k=1}^{K} (a_k - 1 + N_k)$$

I will change the notation for clarity:

$$\lambda = \sum_{k'=1}^{K} (a_{k'} - 1 + N_{k'})$$

Finally, we can get our MAP estimate, our optimal value for θ_k as:

$$\theta_k = \frac{(a_k - 1 + N_k)}{\sum_{k'=1}^{K} (a_{k'} - 1 + N_{k'})}$$

3. **[4pts] Factor Analysis.** *This question is about the EM algorithm. Since some of you will have seen EM in more detail than others before reading week, we have decided to give you the 4 points for free. So you don't need to submit a solution to this part if you don't want to. But we recommend you make an effort anyway, since you probably know enough to solve it, and it will help you practice the course material.*

   In lecture, we covered the EM algorithm applied to mixture of Gaussians models. In this question, we'll look at another interesting example of EM, namely factor analysis. This is a model very similar in spirit to PCA: we have data in a high-dimensional space, and we'd like to summarize it with a lower-dimensional representation. Unlike PCA, we formulate the problem in terms of a probabilistic model. We assume the latent code vector $\mathbf{z}$ is drawn from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and that the observations are drawn from a diagonal covariance Gaussian whose mean is a linear function of $\mathbf{z}$. We'll consider the slightly simplified case of scalar-valued z. The probabilistic model is given by:

$$z \sim \mathcal{N}(0, 1)$$
$$\mathbf{x} \,|\, z \sim \mathcal{N}(z\mathbf{u}, \mathbf{\Sigma}),$$

   where $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. Note that the observation model can be written in terms of coordinates:

$$x_j \,|\, z \sim \mathcal{N}(zu_j, \sigma_j).$$

   We have a set of observations $\{\mathbf{x}^{(i)}\}_{i=1}^N$, and z is a latent variable, analogous to the mixture component in a mixture-of-Gaussians model.

   In this question, we'll derive both the E-step and the M-step for the EM algorithm. If you don't feel like you understand the EM algorithm yet, don't worry; we'll walk you through it, and the question will be mostly mechanical.

   (a) **E-step (2pts).** In this step, our job is to calculate the statistics of the posterior distribution $q(z) = p(z \,|\, \mathbf{x})$ which we'll need for the M-step. In particular, your job is to find formulas for the (univariate) statistics:

$$m = \mathbb{E}[z \,|\, \mathbf{x}] =$$
$$s = \mathbb{E}[z^2 \,|\, \mathbf{x}] =$$

   Tips:

   - Compare the model here with the linear Gaussian model of the Appendix. Note that z here is a scalar, while the Appendix gives the more general formulation where $\mathbf{x}$ and $\mathbf{z}$ are both vectors.
   - Determine $p(z \,|\, \mathbf{x})$. To help you check your work: $p(z \,|\, \mathbf{x})$ is a univariate Gaussian distribution whose mean is a linear function of $\mathbf{x}$, and whose variance does not depend on $\mathbf{x}$.
   - Once you have figured out the mean and variance, that will give you the conditional expectations.

   Solution:

   From the appendix we have:

$$p(\mathbf{z}) = \mathcal{N}\left(\mathbf{z} \,|\, \mu, \Lambda^{-1}\right)$$
$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \,|\, \mathbf{u} \cdot \mu + b, \Sigma + \mathbf{u} \cdot \Lambda^{-1} \cdot \mathbf{u}\right)$$

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{u} \cdot \mathbf{z} + b, \Sigma)$$
$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid C \cdot \left(A^T \cdot \Sigma^{-1} \cdot (x - b) + \Lambda \cdot \mu \right), C\right)$$

The question states that z is drawn from a Gaussian distribution $\mathcal{N}(0, 1)$.

$$b = 0, \mu = 0, \Lambda = 1,$$

$$p(\mathbf{z}) = \mathcal{N}(0, 1)$$
$$p(\mathbf{x}) = \mathcal{N}(0, \Sigma + \mathbf{u} \cdot \mathbf{u})$$
$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{u} \cdot \mathbf{z}, \Sigma)$$
$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid C \cdot \left(\mathbf{u}^T \cdot \Sigma^{-1} \cdot \mathbf{x} \right), C\right)$$

$$C = \left(1 + \mathbf{u}^T \cdot \Sigma^{-1} \cdot \mathbf{u}\right)^{-1}$$

Check: Is the mean a linear function of x? *yes*

Check: Does the variance of the expectation depend on x? *no*

$$mean : \mathbb{E}[\mathbf{z} \mid \mathbf{x}] = C(\mathbf{u}^T \cdot \Sigma^{-1} \cdot \mathbf{x})$$
$$mean : \mathbb{E}[\mathbf{z} \mid \mathbf{x}] = \frac{\mathbf{u}^T \cdot \Sigma^{-1} \cdot \mathbf{x}}{1 + \mathbf{u}^T \cdot \Sigma^{-1} \cdot \mathbf{u}}$$

$$var : \mathbb{E}[\mathbf{z}^2 \mid \mathbf{x}] - \mathbb{E}[\mathbf{z} \mid \mathbf{x}]^2 = C$$
$$var : \mathbb{E}[\mathbf{z}^2 \mid \mathbf{x}] = C + \mathbb{E}[\mathbf{z} \mid \mathbf{x}]^2$$

(b) **M-step (2pts).** In this step, we need to re-estimate the parameters of the model. The parameters are $\mathbf{u}$ and $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. For this part, your job is to derive a formula for $\mathbf{u}_{\text{new}}$ that maximizes the expected log-likelihood, i.e.,

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log p(z^{(i)}, \mathbf{x}^{(i)})].$$

(Recall that $q(z)$ is the distribution computed in part (a).) This is the new estimate obtained by the EM procedure, and will be used again in the next iteration of the E-step. Your answer should be given in terms of the $m^{(i)}$ and $s^{(i)}$ from the previous part. (I.e., you don't need to expand out the formulas for $m^{(i)}$ and $s^{(i)}$ in this step, because if you were implementing this algorithm, you'd use the values $m^{(i)}$ and $s^{(i)}$ that you previously computed.)

Tips:

- Expand $\log p(z^{(i)}, \mathbf{x}^{(i)})$ to $\log p(z^{(i)}) + \log p(\mathbf{x}^{(i)} \mid z^{(i)})$ (log is the natural logarithm).
- Expand out the PDF of the Gaussian distribution.
- Apply linearity of expectation. You should wind up with terms proportional to $\mathbb{E}_{q(z^{(i)})}[z^{(i)}]$ and $\mathbb{E}_{q(z^{(i)})}[[z^{(i)}]^2]$. Replace these expectations with $m^{(i)}$ and $s^{(i)}$. You should get an equation that does not mention $z^{(i)}$.

- In order to find the maximum likelihood parameter $\mathbf{u}_{\text{new}}$, you need to take the derivative with respect to u_j, set it to zero, and solve for $\mathbf{u}_{\text{new}}$.

Solution:
Starting with:

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log p(z^{(i)}, \mathbf{x}^{(i)})]$$

Separate the log using the chain rule.

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q(z^{(i)})}[\log p(x^{(i)} \mid \mathbf{z}^{(i)}) + \log p(\mathbf{z}^{(i)})]$$

The expectation can be distributed across both terms.

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z^{(i)})}[\log p(x^{(i)} \mid \mathbf{z}^{(i)})] + \mathbb{E}_{q(z^{(i)})}[\log p(\mathbf{z}^{(i)})] \Big]$$

Reminder. Gaussian formula is:

$$\mathcal{N}(z, \mu, \sigma) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(z-\mu)^2}{\sigma^2} \right)$$

Since $p(\mathbf{z}^{(i)})$ is Gaussian we can substitute the formula for the Gaussian distribution with unit variance and zero mean as described by the probabilistic model.

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\log p(x^{(i)} \mid \mathbf{z}^{(i)})] + \mathbb{E}_{q(z^{(i)})}[\log \left(2\pi^{-\frac{1}{2}} \cdot \exp(-\frac{1}{2} \cdot (z^{(i)})^2) \right)] \right)$$

The terms can be distributed as:

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\log p(x^{(i)} \mid \mathbf{z}^{(i)})] + \mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \cdot \Big(\log(2\pi) + (z^{(i)})^2 \Big)] \right)$$

We know that $\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{z} \cdot u, \Sigma)$
Focusing on the likelihood,$p(\mathbf{x} \mid \mathbf{z})$, we can compute the expectation of this formula as:

$$\mathbb{E}_{q(z^{(i)})}[\log \left(\frac{1}{\sqrt{\Sigma} \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(x - \mathbf{z} \cdot u)^2}{\Sigma} \right) \right)]$$

$$\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \log \Sigma - \frac{1}{2} \cdot \log(2\pi) - \frac{1}{2} \cdot \frac{(x^{(i)} - \mathbf{z} \cdot u)^2}{\Sigma}]$$

Plugging this expression back into the argmax function yields:
$\arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \Big(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \log \Sigma - \frac{1}{2} \cdot \log(2\pi) - \frac{1}{2} \cdot \frac{(x^{(i)} - \mathbf{z} \cdot u)^2}{\Sigma}] + \mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \cdot \Big(\log(2\pi) + (z^{(i)})^2 \Big)] \Big)$
To maximise this function we can acknowledge and ignore constant values.

$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \log \Sigma - C_1 - \frac{1}{2} \cdot \frac{(x^{(i)} - \mathbf{z} \cdot u)^2}{\Sigma} - C_2 - \frac{1}{2} \cdot (z^{(i)})^2] \right)$
Now we take the partial derivative with respect to $\mathbf{u}$.

$$\frac{\partial}{\partial \mathbf{u}} \cdot f(\mathbf{u}) = \frac{\partial}{\partial \mathbf{u}} \cdot \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \log \Sigma - C_1 - \frac{1}{2} \cdot \frac{(x^{(i)} - \mathbf{z} \cdot \mathbf{u})^2}{\Sigma} - C_2 - \frac{1}{2} \cdot (z^{(i)})^2] \right)$$

$$\frac{\partial}{\partial \mathbf{u}} \cdot f(\mathbf{u}) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[0 - 0 - \frac{1}{2} \cdot (x^{(i)} - \mathbf{z} \cdot \mathbf{u})^T \cdot \Sigma^{-1} \cdot \mathbf{z} - 0 - 0] \right)$$

Setting the derivative to 0 allows us to find the optimal point.

$$0 = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \cdot (x^{(i)} - \mathbf{z} \cdot /bu)^T \cdot \Sigma^{-1} \cdot \mathbf{z}] \right)$$

$$0 = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \cdot x^{(i)} \cdot \Sigma^{-1} \cdot \mathbf{z}] \right) - \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2} \cdot \mathbf{z} \cdot \mathbf{u} \cdot \Sigma^{-1} \cdot \mathbf{z}] \right)$$

$$\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\cdot x^{(i)} \cdot \mathbf{z}] \right) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z} \cdot \mathbf{u} \cdot \mathbf{z}] \right)$$

$$\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(i)} \right) = \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\cdot \mathbf{z} \cdot \mathbf{z}] \right) \cdot \mathbf{u}$$

$$\mathbf{u}_{new} = \frac{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(i)} \right)}{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z} \cdot \mathbf{z}] \right)}$$

Now we must apply the linearity of expectation to eliminate terms referring to $\mathbf{z}$ as noted in the tips.
recall that :

$$m : \mathbb{E}[\mathbf{z} \,|\, \mathbf{x}]C$$

$$s : \mathbb{E}[\mathbf{z}^2 \,|\, \mathbf{x}]$$

since z and z^2 are completely dependant, the information provided by conditioning on x does not change the proportion of the expected value of z and z^2. Thus:

$$\mathbf{u}_{new} = \frac{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(i)} \right)}{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \right)} = \frac{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z} \,|\, \mathbf{x}] \cdot x^{(i)} \right)}{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2 \,|\, \mathbf{x}] \right)} = \frac{\sum_{i=1}^{N} \left(m^{(i)} \cdot x^{(i)} \right)}{\sum_{i=1}^{N} s^{(i)}}$$

(c) **M-step, cont'd (optional)** Find the M-step update for the observation variances $\{\sigma_j\}_{j=1}^D$. This can be done in a similar way to part (b).

Solution: We want to maximise our expectation of the variance given our observed evidence. More specifically, we want to maximise to following equation:

$$\frac{\partial}{\partial \Sigma} \cdot f(\Sigma) = \frac{\partial}{\partial \Sigma} \cdot \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[-\frac{1}{2}\log \Sigma - C_1 - \frac{1}{2} \cdot \frac{(x^{(i)} - \mathbf{z} \cdot \mathbf{u})^2}{\Sigma} - C_2 - \frac{1}{2} \cdot (z^{(i)})^2] \right)$$

Setting the derivative to zero gives us:

$$\frac{\partial}{\partial \Sigma^{-1}} = 0 = \Sigma^T - \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[(x^{(n)} - \mathbf{u} \cdot \mathbf{z}) \cdot (x^{(n)} - \mathbf{u} \cdot \mathbf{z})^T] \right)$$

The terms can be expanded and distributed:

$$0 = \Sigma^T - \frac{1}{N} \sum_{n=1}^{N} \Big(\ x^{(n)} \cdot (x^{(n)})^T - 2x^{(n)} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}] + \mathbf{u}_{new} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \Big)$$

Rearranging the terms:

$$\Sigma^T = \frac{1}{N} \sum_{n=1}^{N} \Big(\ x^{(n)} \cdot (x^{(n)})^T - 2x^{(n)} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}] + \mathbf{u}_{new} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \Big)$$

If covariance is diagonal (i.e. all features independent)

$$\Sigma_{new} = \frac{1}{N} \sum_{n=1}^{N} \Big(\ x^{(n)} \cdot (x^{(n)})^T - 2x^{(n)} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}] + \mathbf{u}_{new} \cdot \mathbf{u}_{new}^T \cdot \mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \Big)$$

Distribute the summations

$$\Sigma_{new} = \frac{1}{N} \sum_{n=1}^{N} \left(x^{(n)} \cdot (x^{(n)})^T \right) - 2 \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(n)} \right) \cdot \mathbf{u}_{new}^T + \mathbf{u}_{new} \cdot \mathbf{u}_{new}^T \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \right)$$

Substitute $\mathbf{u}_{new}^T$.

$$\Sigma_{new} = \frac{1}{N} \sum_{n=1}^{N} \left(x^{(n)} \cdot (x^{(n)})^T \right) - 2 \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(n)} \right) \cdot \frac{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(i)} \right)^T}{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}^2] \right)} + \frac{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot \right.}{\sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \right.}$$

These terms can be collected again into the update means:

$$\Sigma_{new} = \frac{1}{N} \sum_{n=1}^{N} \left(x^{(n)} \cdot (x^{(n)})^T \right) - 2 \cdot \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(n)})}[\mathbf{z}] \cdot x^{(n)} \right)^T \cdot \mathbf{u}_{new} + \frac{1}{N} \mathbf{u}_{new} \sum_{i=1}^{N} \left(\mathbb{E}_{q(z^{(i)})}[\mathbf{z}] \cdot x^{(i)} \right)^T$$

$$\Sigma_{new} = \frac{1}{N} \sum_{n=1}^{N} \left(x^{(n)} \cdot (x^{(n)})^T \right) - \frac{1}{N} \sum_{n=1}^{N} \left(\mathbb{E}_{q(z^{(n)})}[\mathbf{z}] \cdot x^{(n)} \right)^T \cdot \mathbf{u}_{new}$$

Appendix: Some Properties of Conditional Gaussians

Consider a multivariate Gaussian random variable $\mathbf{z}$ with the mean μ and the covariance matrix Λ^{-1} (Λ is the inverse of the covariance matrix and is called the precision matrix). We denote this by

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{z} \,|\, \mu, \Lambda^{-1}).$$

Now consider another Gaussian random variable $\mathbf{x}$, whose mean is an affine function of $\mathbf{z}$ (in the form to be clear soon), and its covariance L^{-1} is independent of $\mathbf{z}$. The conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is

$$p(\mathbf{x} \,|\, \mathbf{z}) = \mathcal{N}(\mathbf{x} \,|\, A\mathbf{z} + b, L^{-1}).$$

Here the matrix A and the vector b are of appropriate dimensions.

In some problems, we are interested in knowing the distribution of $\mathbf{z}$ given $\mathbf{x}$, or the marginal distribution of $\mathbf{x}$. One can apply Bayes' rule to find the conditional distribution $p(\mathbf{z} \,|\, \mathbf{x})$. After some calculations, we can obtain the following useful formulae:

$$p(\mathbf{x}) = \mathcal{N}\left(x \,|\, A\mu + b, L^{-1} + A\Lambda^{-1}A^\top\right)$$
$$p(\mathbf{z} \,|\, \mathbf{x}) = \mathcal{N}\left(x \,|\, C(A^\top L(x - b) + \Lambda\mu), C\right)$$

with

$$C = (\Lambda + A^\top L A)^{-1}.$$