SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #2

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *October 7th, 2020*

Question 1

(a) **[True or False]** If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

(b) We consider a partitioning of the components of x into three groups x_a, x_b, and x_c, with a corresponding partitioning of the mean vector μ and of the covariance matrix $\sum$ in the form

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}.$$

Find an expression for the conditional distribution $p(x_a|x_b)$ in which x_c has been marginalized out.

Answer. (a) True.
(b) We first of all take the joint distribution $p(x_a, x_b, x_c)$ and marginalize to obtain the distribution $p(x_a, x_b)$. According to the properties of marginal Gaussian distributions, this is again a Gaussian distribution with mean and covariance given by

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

From the results of conditioanl Gaussiandistributions, the distribution $p(x_a, |x_b)$ is then Gaussian with mean and covariance given by

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b),$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.$$

Question 2

Consider a joint distribution over the variable

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

whose mean and covariance are given by

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mu \\ \mathbf{A}\mu + \mathbf{b} \end{pmatrix}, \quad \mathrm{cov}[\mathbf{z}] = \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix}.$$

(a) Show that the marginal distribution $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda}^{-1})$.

(b) Show that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is given by $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$.

Answer. (a) For the marginal distribution $p(\mathbf{x})$, we can see from the results of marginal Gaussian distributions that the mean is given by the upper partition of $\mathbb{E}[\mathbf{z}]$ which is simply μ. Similarly, the covariance is given by the top left partition of $cov[\mathbf{z}]$ which is $\mathbf{\Lambda}^{-1}$.
(b) According to the expression for the mean of conditional Gaussian distribution $p(\mathbf{x}_a|\mathbf{x}_b)$, which is

$$\mu_{a|b} = \mu_a + \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \mu_b),$$

we obtain

$$\mu_{\mathbf{y}|\mathbf{x}} = \mathbf{A}\mu + \mathbf{b} + \mathbf{A}\mathbf{\Lambda}_{-1}\mathbf{\Lambda}(\mathbf{x} - \mu) = \mathbf{A}\mathbf{x} + \mathbf{b}.$$

Similarly, according to the expression for covariance

$$\mathbf{\Sigma}_{a|b} = \mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba},$$

we have

$$cov[\mathbf{y}|\mathbf{x}] = \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathbf{T}} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^{\mathbf{T}} = \mathbf{L}^{-1}$$

Question 3

Show that the covariance matrix Σ that maximizes the log likelihood function is given by the sample covariance

$$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^{\mathrm{T}}.$$

Is the final result symmetric and positive definite (provided the sample covariance is nonsingular)?

Hints.

(*a*) To find the maximum likelihood solution for the covariance matrix of a multivariate Gaussian, we need to maximize the log likelihood function with respect to Σ. The log likelihood function is given by

$$\ln p(\mathbf{X}|\mu,\Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n-\mu).$$

(*b*) The derivative of the inverse of a matrix can be expressed as

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$$

We have the following properties

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{A}) = \mathbf{I}, \quad \frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A}^{-\mathbf{1}})^{\mathrm{T}}.$$

Answer. Differentiating the log likelihood function with respect to Σ we obtain two terms:

$$-\frac{N}{2}\frac{\partial}{\partial \mathbf{\Sigma}}ln|\mathbf{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{n-1}^{N}(\mathbf{x}_n-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\mu).$$

For the first term, according to the third equation in hint(2), we can get

$$-\frac{N}{2}\frac{\partial}{\partial \mathbf{\Sigma}}ln|\mathbf{\Sigma}| = -\frac{N}{2}(\mathbf{\Sigma}^{-1})^T = -\frac{N}{2}\mathbf{\Sigma}^{-1}.$$

For the second term, we first re-write the sum

$$\sum_{n-1}^{N}(\mathbf{x}_n-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\mu) = N\mathrm{Tr}[\mathbf{\Sigma}^{-1}\mathbf{S}],$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T.$$

Using together with the first equation in hint(2), in which $x = \sum_{ij}$ (element(i,j) in Σ), and properties of the trace we get

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\mu) &= N\frac{\partial}{\partial \Sigma_{ij}}\mathrm{Tr}[\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}] \\
&= N\mathrm{Tr}[\frac{\partial}{\partial \Sigma_{ij}}\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}] \\
&= -N\mathrm{Tr}[\mathbf{\Sigma}^{-1}\frac{\partial}{\partial \Sigma_{ij}}\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}] \\
&= -N\mathrm{Tr}[\frac{\partial \mathbf{\Sigma}}{\partial \Sigma_{ij}}\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}\mathbf{\Sigma}^{-1}] \\
&= -N(\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}\mathbf{\Sigma}^{-\mathbf{1}})_{ij},
\end{aligned}$$

Where we have used the second equation. Note that in the last step we have ignored the fact that $\Sigma_{ij} = \Sigma_{ji}$, so that $\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}$ has a 1 in position (i, j) only an 0 everywhere else. Treating this result as valid neverthe less, we get

$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \mu) = \frac{N}{2}\boldsymbol{\Sigma^{-1}S\Sigma^{-1}}.$$

Combining the derivatives of the two terms and setting the result to zero, we obtain

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma^{-1}S\Sigma^{-1}}.$$

Re-arrangement then yields $\boldsymbol{\Sigma} = \mathbf{S}$.
Yes.

Question 4

(*a*) Derive an expression for the sequential estimation of the variance of a univariate Gaussian distribution, by starting with the maximum likelihood expression

$$\sigma^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

(*b*) Derive an expression for the sequential estimation of the covariance of a multivariate Gaussian distribution, by starting with the maximum likelihood expression

$$\Sigma_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\text{ML}})(\mathbf{x}_n - \mu_{\text{ML}})^{\text{T}}.$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

Hints.

(*a*) Consider the result $\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ for the maximum likelihood estimator of the mean μ_{ML}, which we will denote by $\mu_{\text{ML}}^{(N)}$ when it is based on N observations. If we dissect out the contribution from the final data point $\mathbf{x}_N$, we obtain

$$\mu_{\text{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{\text{ML}}^{(N-1)}$$

$$= \mu_{\text{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \mu_{\text{ML}}^{(N-1)}).$$

(*b*) Robbins-Monro for maximum likelihood

$$\theta^{(N)} = \theta^{(N-1)} + a_{(N-1)}\frac{\partial}{\partial \theta^{(N-1)}}\ln p(x_N|\theta^{(N-1)}).$$

Answer. (a) Consider the expression for $\sigma^2(N)$ and separate out the contribution from observation x_N to give

$$\begin{aligned}
\sigma^2_{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(x_n-\mu)^2 + \frac{(x_N-\mu)^2}{N} \\
&= \frac{N-1}{N}\sigma^2_{(N-1)} + \frac{(x_N-\mu)^2}{N} \\
&= \sigma^2_{(N-1)} - \frac{1}{N}\sigma^2_{(N-1)} + \frac{(x_N-\mu)^2}{N} \\
&= \sigma^2_{(N-1)} + \frac{1}{N}\{(x_n-\mu)^2 - \sigma^2_{(N-1)}\}.
\end{aligned}$$

If we substitute the expression for a Gaussian distribution into the result of hint(2), we obtain

$$\begin{aligned}
\sigma^2_{(N)} &= \sigma^2_{(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\{-\frac{1}{2}\ln\sigma^2_{(N-1)} - \frac{(x_N-\mu)^2}{2\sigma^2_{(N-1)}}\} \\
&= \sigma^2_{(N-1)} + a_{N-1}\{-\frac{1}{2\sigma^2_{N-1}} + \frac{(x_N-\mu)^2}{2\sigma^4_{(N-1)}}\} \\
&= \sigma^2_{(N-1)} + \frac{a_{N-1}}{2\sigma^4_{(N-1)}}\{(x_N-\mu)^2 - \sigma^2_{(N-1)}\}.
\end{aligned}$$

Comparison of these two results allows us to identify

$$a_{N-1} = \frac{2\sigma^4_{(N-1)}}{N}.$$

(b)

$$\begin{aligned}
\boldsymbol{\Sigma}^{(N)}_{ML} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T + \frac{1}{N}(\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}^{(N-1)}_{ML} + \frac{1}{N}(\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T \\
&= \boldsymbol{\Sigma}^{(N-1)}_{ML} + \frac{1}{N}((\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T - \boldsymbol{\Sigma}^{(N-1)}_{ML}).
\end{aligned}$$

From Answer 3, we know that

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\Sigma}^{(N-1)}_{ML}}\ln p(\mathbf{x}_N|\mu,\boldsymbol{\Sigma}^{(N-1)}_{ML}) \\
&= \frac{1}{2}(\boldsymbol{\Sigma}^{(N-1)}_{ML})^{-1}((\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T - \boldsymbol{\Sigma}^{(N-1)}_{ML})(\boldsymbol{\Sigma}^{(N-1)}_{ML})^{-1} \\
&= \frac{1}{2}(\boldsymbol{\Sigma}^{(N-1)}_{ML})^{-2}((\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T - \boldsymbol{\Sigma}^{(N-1)}_{ML}),
\end{aligned}$$

where we have used the assumption that $\mathbf{\Sigma}_{ML}^{(N-1)}$, and hence $(\mathbf{\Sigma}_{ML}^{(N-1)})^{-1}$ is diagonal. If we substitute this into the multivariate form of hint(2), we get

$$\mathbf{\Sigma}_{ML}^{(N)} = \mathbf{\Sigma}_{ML}^{(N-1)} + \mathbf{A}_{N-1}\frac{1}{2}(\mathbf{\Sigma}_{ML}^{(N-1)})^{-2}((\mathbf{x}_n-\mu)(\mathbf{x}_n-\mu)^T - \mathbf{\Sigma}_{ML}^{(N-1)}),$$

where $\mathbf{A}_{N-1}$ is a matrix of coefficients corresponding to a_{N-1} in hint(2). By comparing these two results, we see

$$\mathbf{A}_{N-1} = \frac{2}{N}(\mathbf{\Sigma}_{ML}^{(N-1)})^2.$$

Question 5

Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $N(x|\mu,\Sigma)$ in which th4e covariance Σ is known and for which we wish to infer the mean μ from a set of observations $\mathbf{X} = \{x_1, x_2, ..., x_N\}$. Given a prior distribution $p(\mu) = N(\mu|\mu_0,\Sigma_0)$, find the corresponding posterior distribution $p(\mu|\mathbf{X})$.

Answer. The posterior distribution is proportional to the product of the prior and the likelihood function

$$p(\mu|\mathbf{X}) \propto p(\mu)\prod_{n=1}^{N} p(\mathbf{x}_n|\mu,\mathbf{\Sigma}).$$

Thus the posterior is proportional to an exponential of a quadratic form inμ given by

$$\begin{aligned}
&-\frac{1}{2}(\mu-\mu_0)^T\mathbf{\Sigma}_0^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\mu)\\
&= -\frac{1}{2}\mu^T(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})\mu + \mu^T(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{n=1}^{N}\mathbf{x}_n) + const,
\end{aligned}$$

where 'const' denotes terms independent of μ. The mean and covariance of the posterior distribution are given by

$$\begin{aligned}
\mu_N &= (\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})^{-1}(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}N\mu_{ML})\\
\mathbf{\Sigma}_N^{-1} &= \mathbf{\Sigma}_0^{-1} + N\mathbf{\Sigma}^{-1},
\end{aligned}$$

where μ_{ML} is the maximum likelihood solution for the mean given by

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$