Mohammadsadegh Abolhasani
u1416052
HW1 CS6190
Spring 23

Analytical problems

1.

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, 1 : (1, ..., k), 2 : (k = 1, .., n)$$

$$\mathbf{x}_1 = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_k \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{x}_n \end{bmatrix}$$

$$\mathbf{\Sigma}_{11} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \dots & \mathbf{\Sigma}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{k1} & \dots & \mathbf{\Sigma}_{kk} \end{bmatrix}, \text{ the same is similar for other 3 matrices.}$$

Next, we should employ the properties of matrices. First is the affine property. For this purpose, we have first to define a projection matrix $\mathbf{A}$ that is a $k * n$ matrix.

$$\mathbf{Ax} = \mathcal{N}(\mathbf{Ax}|\mathbf{A}\mu, \mathbf{A\Sigma A}^T)$$

$$\mathbf{A} = \begin{bmatrix} 1 & \dots & 0 & 0 \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \dots & \dots & 0 \end{bmatrix}$$

$$\mathbf{x} = \mathbf{A} = \begin{bmatrix} 1 & \dots & 0 & 0 \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \dots & \dots & 0 \end{bmatrix} * \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_k \\ \mathbf{x}_{k+1} \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_k \end{bmatrix} = \mathbf{x}_1$$

$$\mathbf{A\Sigma A}^T = \begin{bmatrix} 1 & \dots & 0 & 0 \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \dots & \dots & 0 \end{bmatrix} * \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} * \left(\begin{bmatrix} 1 & \dots & 0 & 0 \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \dots & \dots & 0 \end{bmatrix} \right)^T =$$

$$= \begin{bmatrix} \mathbf{\Sigma}_{11} & \dots & \mathbf{\Sigma}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{k1} & \dots & \mathbf{\Sigma}_{kk} \end{bmatrix} = \mathbf{\Sigma}_{11} -> \mathbf{x}_1 = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$$

2. A characteristic function is a function that accurately captures the probability distribution of a random variable.

$$\phi(t) = \mathbb{E}[e^{itX}]$$

where t is a real number and I is the fictitious unit.
From the last question, we check the property for the Gaussian distribution. Based on that, we know:

$$\mathbf{Ax} = \mathcal{N}(\mathbf{Ax}|\mathbf{A}\mu, \mathbf{A\Sigma A}^T) -> \mathbf{Ax} + \mathbf{b} + \mathbf{z} \sim N(\mathbf{Ax} + \mathbf{b}, \mathbf{A\Sigma A}^T)$$

We now compile the characteristic function for $\mathbf{x}$, which follows the normal distribution.

$$\phi(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{X}}] = e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\mathbf{\Sigma}\boldsymbol{t}}$$

Therefore, for every $(A, \mathbf{b})$, and $\mathbf{y}$ (as mentioned in the question) we should have:

$$\mathbb{E}(e^{it^T Y}) = e^{it^T \mathbf{b}} E(e^{it^A \mathbf{x}})$$

$$\mathbb{E}(e^{it^T Y}) = e^{it^T \mathbf{b}} e^{it^T A\mu - t^T A\mathbf{\Sigma} A^T t/2} = e^{it^T \mathbf{\Sigma} t^T A\mu - t^T A\mathbf{\Sigma} A^T t/2}$$

The equation above is the characteristic function for $Y = A\mathbf{x} + \mathbf{b} \sim N(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T)$ which proves Y follows the Gaussian distribution. In addition, while using the characteristic function we have to show that combining two Gaussian R.V.s will also follow Gaussian. By searching for the characteristics function and its properties for this case I found:

$$\phi_{X+Y}(t) = \phi_X(\mathbf{t})\phi_Y(\mathbf{t})$$

And we know that $\mathbf{x}$ and $\mathbf{y}$ both follow the normal distribution.

$$\mathbf{x} \sim N(\mu_1, \sigma_1), \mathbf{y} \sim N(\mu_2, \sigma_2)$$

from the last part of the proof we have:

$$\phi_X(t) = e^{i\mu_1 t - \frac{1}{2}\sigma_1^2 t^2}, \phi_Y(t) = e^{i\mu_2 t - \frac{1}{2}\sigma_2^2 t^2}$$

$$\phi_Z(t) = \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = e^{i(\mu_1+\mu_2)t - \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}$$

The above characteristics function clearly shows a Gaussian distribution that has a mean of $(\mu_1 + \mu_2)$ and variance of $(\sigma_1^2 + \sigma_2^2)$. This shows that the combination of two Gaussian R.V.s is still Gaussian. Back to the question, for $\mathbf{z}$ we have the mean to be equal to zero. Therefore, putting it to the above equation will not change it.

3.

$$H[\mathbf{x}] = -\mathbf{\Sigma} p(\mathbf{x}) log(p(\mathbf{x}))$$

$$= -\int \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) log(\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}))dx = -\mathbb{E}[log(\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}))] \ (1)$$

$$log(\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})) = -\frac{d}{2}log(2\pi) - \frac{1}{2}log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu) \ (2)$$

Considering (1) and (2) together, and keeping in mind that we can rearrange quadratics from right to left step by step, we can combine them and have:

$$H[\mathbf{x}] = \frac{d}{2}log(2\pi) + \frac{1}{2}log|\mathbf{\Sigma}| + \frac{1}{2}\mathbb{E}[(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu)]$$

$$\begin{aligned}
\mathrm{E}\left[(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right] &= \mathrm{E}\left[\mathrm{tr}\left((\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right)\right] \\
&= \mathrm{E}\left[\mathrm{tr}\left(\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\right)\right] \\
&= \mathrm{tr}\left(\Sigma^{-1}\mathrm{E}\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathrm{T}}\right]\right) \\
&= \mathrm{tr}\left(\Sigma^{-1}\Sigma\right) \\
&= \mathrm{tr}\left(I_n\right) \\
&= d
\end{aligned}$$

$$-> \frac{1}{2}\mathbb{E}[(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu)] = \frac{d}{2}$$

$$-> H[\mathbf{x}] = \frac{d}{2}log(2\pi) + \frac{1}{2}log|\mathbf{\Sigma}| + \frac{d}{2} = \frac{d}{2}(1 + log(2\pi)) + \frac{1}{2}log|\mathbf{\Sigma}|$$

4.

$$KL(q||p) = -\int (\mathbf{x}) ln(\frac{q(\mathbf{x})}{p(\mathbf{x})}) d\mathbf{x} = \int q(\mathbf{x}) lnq(\mathbf{x}) d\mathbf{x} - \int p(\mathbf{x}) lnp(\mathbf{x}) d\mathbf{x} = -H_q[\mathbf{x}] - \mathbb{E}_q[lnp(\mathbf{x})]$$

We should use the result from question (3) to solve this problem because this form represents our finding in that question.

$$KL(q||p) = -\frac{1}{2} ln|\Lambda| - \frac{d}{2}(1 + ln(2\pi)) - \mathbb{E}_q[ln(p(\mathbf{x}))]$$

From question (3) we have:

$$H[\mathbf{x}] = \frac{d}{2}(1 + log(2\pi)) + \frac{1}{2} log|\mathbf{\Sigma}|$$

Now we will rewrite the form for this question:

$$KL(q||p) = -\frac{d}{2}(1 + log(2\pi)) - \frac{1}{2} log|\Lambda| - \mathbb{E}_q[ln((2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu))]$$

$$= -\frac{d}{2}(1 + log(2\pi)) - \frac{1}{2} log|\Lambda| + \frac{d}{2}(ln(2\pi)) + \frac{1}{2} log|\mathbf{\Sigma}| + \mathbb{E}_q[Tr((\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu))]$$

We use eq. 16 of Matrix Cookbook

$$= -\frac{1}{2} ln\frac{|\mathbf{\Sigma}|}{|\Lambda|} - \frac{d}{2} + \frac{1}{2} Tr[\mathbb{E}_q[\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)^T (\mathbf{x} - \mu)]] \ (1)$$

I asked the professor for help with the next step, where he advised me to add and negate $\mathbf{m}$ so I can introduce the parameters of $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \Lambda)$ to the equation and get to the final form.

$$\frac{1}{2} Tr[\mathbb{E}_q[\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)^T (\mathbf{x} - \mu)]] = \frac{1}{2} Tr[\mathbf{\Sigma}^{-1} \mathbb{E}_q[(\mathbf{x} - \mathbf{m} + \mathbf{m} - \mu)(\mathbf{x} - \mathbf{m} + \mathbf{m} - \mu)^T]]$$

$$(1) : -\frac{1}{2} ln\frac{|\mathbf{\Sigma}|}{|\Lambda|} - \frac{d}{2} + \frac{1}{2} Tr[\mathbf{\Sigma}^{-1}[\Lambda + 2(\mathbf{m} - \mathbf{m})(\mathbf{m} - \mu)^T + (\mathbf{m} - \mu)(\mathbf{m} - \mu)^T]]$$

$$= -\frac{1}{2} ln\frac{|\mathbf{\Sigma}|}{|\Lambda|} - \frac{d}{2} + \frac{1}{2} Tr[\mathbf{\Sigma}^{-1}[\Lambda + (\mathbf{m} - \mu)(\mathbf{m} - \mu)^T]]$$

$$= -\frac{1}{2} ln\frac{|\mathbf{\Sigma}|}{|\Lambda|} - \frac{d}{2} + \frac{1}{2} Tr[\mathbf{\Sigma}^{-1}\Lambda] + \frac{1}{2}(\mathbf{m} - \mu)\mathbf{\Sigma}^{-1}(\mathbf{m} - \mu)^T$$

5.

$$Z(\eta) = \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) d\mathbf{x}$$

$$\frac{\partial \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{1}{Z(\boldsymbol{\eta})} \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) \mathbf{u}(\mathbf{x}) d\mathbf{x}$$

$$\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = \frac{\partial}{\partial \boldsymbol{\eta}} \frac{1}{Z(\boldsymbol{\eta})} \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) \mathbf{u}(\mathbf{x}) d\mathbf{x}$$

$$= -\frac{1}{z(\eta)^2} \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) \mathbf{u}(\mathbf{x}) d\mathbf{x} \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) \mathbf{u}(\mathbf{x})^T d\mathbf{x} + \frac{1}{Z(\boldsymbol{\eta})} \int h(\mathbf{x}) exp(\mathbf{u}(\mathbf{x})^T \eta) \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^T d\mathbf{x} \ (1)$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{h(\mathbf{x})}{Z(\boldsymbol{\eta})} \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)., \int p(\mathbf{x}|\boldsymbol{\eta}) = 1 \ (2)$$

$$(1), (2) => \frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2} = \int p(\mathbf{x}|\boldsymbol{\eta}) u(\mathbf{x}) d\mathbf{x} \int p(\mathbf{x}|\boldsymbol{\eta}) u(\mathbf{x})^T d\mathbf{x} + \int p(\mathbf{x}|\boldsymbol{\eta}) u(\mathbf{x}) u(\mathbf{x})^T d\mathbf{x}$$

$$= -\mathbb{E}(u(\mathbf{x}))\mathbb{E}(u(\mathbf{x})^T) + \mathbb{E}(u(\mathbf{x})u(\mathbf{x})^T) = \text{cov}(u(\mathbf{x}))$$

6. From the question 5 we proved that $\frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}$ equals $\text{cov}(\mathbf{u}(\mathbf{x}))$, where covariance is an always non-negative variable. Therefore, the second derivative of $\log Z(\boldsymbol{\eta})$ is always non-negative. Back to the convex properties, we knew that a convex function must be derivable twice, and the second derivative should never fall below zero. Therefore, we can see that the required properties have been met here and $\log Z(\boldsymbol{\eta})$ is convex.

7.
$$I(\mathbf{x},\mathbf{y}) = KL(p(\mathbf{x},\mathbf{y})||p(x)p(y)) = -\int\int p(\mathbf{x},\mathbf{y}) Ln(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}) d\mathbf{x}d\mathbf{y}, p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$$
$$= \int p(\mathbf{x},\mathbf{y}) Ln(p(\mathbf{x}|\mathbf{y})) - \int p(\mathbf{x},\mathbf{y}) Ln(p(\mathbf{x})) d\mathbf{x}d\mathbf{y}$$
$$= -H[\mathbf{x}|\mathbf{y}] + H[\mathbf{x}]$$

8.

- Dirichlet distribution

$$p(\mu_1,\ldots,\mu_k|\mathbf{a}_1,\ldots,\mathbf{a}_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \mu_i^{\alpha_i - 1}$$
$$\frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} = \frac{1}{B(\mathbf{a})}$$
$$p(\mathbf{x},\mathbf{a}) = \frac{1}{B(\mathbf{a})} \prod_{i=1}^k \mu_i^{\alpha_i - 1}$$
$$= \frac{1}{B(\mathbf{a})} \frac{\prod exp(ln(\mathbf{x}_i)\mathbf{a}_i)}{\prod \mathbf{x}_i}$$
$$= \frac{1}{B(\mathbf{a})} \frac{exp(\Sigma ln(\mathbf{x}_i)\mathbf{a}_i)}{\prod \mathbf{x}_i}$$
$$\frac{1}{\prod \mathbf{x}_i B(\mathbf{a})} exp(\Sigma ln(\mathbf{x}_i)\mathbf{a}_i)$$
$$\eta = [-\mathbf{a}_1,\ldots,-\mathbf{a}_K]^T$$
$$\mathbf{u}(\mathbf{x}) = [ln\mathbf{x}_1,\ldots,ln\mathbf{x}_K]^T$$
$$h(\mathbf{x}) = \frac{1}{\prod \mathbf{x}_i}$$
$$\mathbf{Z}(\eta) = \mathbf{B}(\mathbf{a})$$

- Gamma distribution

$$p(x) = \frac{1}{\beta^{-\alpha}\Gamma(\alpha)} \mathbf{x}^{\alpha-1} \exp(-\beta\mathbf{x})$$
$$= \frac{\frac{1}{x}}{\beta^{-\alpha}\Gamma(\alpha)} \mathbf{x}^{\alpha} \exp(-\beta\mathbf{x})$$
$$= \frac{\frac{1}{x}}{\beta^{-\alpha}\Gamma(\alpha)} exp(\mathbf{a}ln\mathbf{x}) \exp(-\beta\mathbf{x})$$
$$= \frac{\frac{1}{x}}{\beta^{-\alpha}\Gamma(\alpha)} exp(\mathbf{a}ln\mathbf{x} - \beta\mathbf{x})$$
$$\eta = [\mathbf{a}\ \beta]$$
$$\mathbf{u}(\mathbf{x}) = [-ln\mathbf{x}\ \mathbf{x}]$$
$$h(\mathbf{x}) = \frac{1}{\mathbf{x}}$$
$$\mathbf{Z}(\eta) = \beta^{-\alpha}\Gamma(\alpha)$$

- Wishart distribution

$$p(\Lambda|d,\mathbf{W}) = \frac{|\Lambda|^{(v-d-1)/2}\exp -\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\Lambda)}{2^{dv/2}|\mathbf{W}|^{v/2}\Gamma_d(\frac{v}{2})}$$

$$= \frac{|\Lambda|^{(v-1)/2}}{2^{dv/2}|\mathbf{W}|^{v/2}\Gamma_d(\frac{v}{2})}|\Lambda|^{(-d)/2}\exp -\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\Lambda)$$

$$= \frac{|\Lambda|^{(v-1)/2}}{2^{dv/2}|\mathbf{W}|^{v/2}\Gamma_d(\frac{v}{2})}|\Lambda|^{(-d)/2}\exp(-\frac{d}{2}ln|\mathbf{x}| - \frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\Lambda))$$

$$\eta = [d/2\ \mathbf{W}^{-1}]$$

$$\mathbf{u}(\mathbf{x}) = [ln|\mathbf{x}|\ \ \frac{\mathbf{x}}{2}]$$

$$h(x) = \Lambda|^{(v-1)/2}$$

$$\mathbf{Z}(\eta) = 2^{dv/2}|\mathbf{W}|^{v/2}\Gamma_d(\frac{v}{2})$$

9. An exponential family is in the following general term:

$$p(\mathbf{x}|\eta) = h(x)exp(\eta^T t(\mathbf{x}) - a(\eta))$$

Due to the fact that the student-t distribution is created by combining a normal distribution with a gamma-distributed accuracy prior, it is not regarded as belonging to the exponential family. h(x) is a normalizing constant, η is a collection of parameters, T(x) is an adequate statistic, and A(η) is a log normalizing function. Distributions that can be expressed in the form mentioned above are the only ones that fall under the umbrella of the exponential family. Since the student-t distribution is created from a combination of other distributions, it cannot be represented in this way because this goes against one of the requirements for membership in the exponential family.

10. No, a combination of Gaussian distributions is not an exponential. Like problem 9, we can see that we can't add these two Gaussian distributions together and make the general format of exponential family (the format in the exp will not match with different sets of parameters from different Gaussian distributions adding to a single exp function).

11.

(a)

$$\mathbf{F}(\eta) = -\mathbb{E}_{p(\mathbf{x}|\eta)}[\frac{\partial^2 \log(p(\mathbf{x}|\eta))}{\partial\eta^2}]$$

$$p(\mathbf{x}|\eta) = \frac{h(\mathbf{x})}{Z(\eta)}exp(u(\mathbf{x})^T\eta)$$

$$lnp(\mathbf{x}|\eta) = -lnZ(\eta) + u(\mathbf{x})^T\eta + lnh(\mathbf{x})$$

$$\frac{\partial \log(p(\mathbf{x}|\eta))}{\partial\eta} = -\frac{\partial \ln(Z(\eta))}{\partial\eta} + u(\mathbf{x})$$

$$\frac{\partial^2 \log(p(\mathbf{x}|\eta))}{\partial\eta^2} = -\frac{\partial^2 \ln(Z(\eta))}{\partial^2\eta}$$

$$\frac{\partial^2 \log(p(\mathbf{x}|\eta))}{\partial\eta^2} = \mathbb{E}[\frac{\partial^2 \ln(Z(\eta))}{\partial^2\eta}]$$

Back to the problem (5), we proved that this is equivalent to cov($u(\mathbf{x})$)

(b) Assume $h(\mathbf{x})$ is 1

$$p(\mathbf{x}|\eta) = exp(u(\mathbf{x})^T\eta - log\mathbf{Z}(\eta)) = \frac{1}{\mathbf{Z}(\eta)}exp(\mathbf{u}(\mathbf{x})^T\eta)$$

From question 5, and the last part of this question we have:

$$\mathbf{F}(\eta) = \text{cov}(u(\mathbf{x})) = \frac{\partial^2 \log Z(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}$$

$$\int p(\mathbf{x}|\boldsymbol{\eta})u(\mathbf{x})d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

$$\int \frac{1}{\mathbf{Z}(\eta)}exp(\mathbf{u}(\mathbf{x})^T\eta)u(\mathbf{x})d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

$$\frac{\partial \mathbb{E}[\mathbf{u}(\mathbf{x})]}{\partial \eta} = \frac{\partial}{\partial \eta}\frac{1}{\mathbf{Z}(\eta)}\int exp(\mathbf{u}(\mathbf{x})^T\eta)u(\mathbf{x})d\mathbf{x}$$

$$= -\frac{1}{z(\eta)^2}\int exp(\mathbf{u}(\mathbf{x})^T\eta)\mathbf{u}(\mathbf{x})d\mathbf{x}\int exp(\mathbf{u}(\mathbf{x})^T\eta)\mathbf{u}(\mathbf{x})^T d\mathbf{x} + \frac{1}{Z(\boldsymbol{\eta})}\int exp(\mathbf{u}(\mathbf{x})^T\eta)\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^T d\mathbf{x}$$

This is the form that we had in question 5, we know we have $\int p(\mathbf{x}|\eta) = 1$ and $h(\mathbf{x}) = 1$, therefore if we substitute these into the equation, we see that it is equal to $\text{cov}(u(\mathbf{x}))$

$$\frac{\partial \theta}{\partial \eta} = \frac{\partial \mathbb{E}[\mathbf{u}(\mathbf{x})]}{\partial \eta} = \text{cov}(u(\mathbf{x})) = \mathbf{F}(\eta)$$

(c)

$$\mathbf{F}(\eta) = -\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial^2 \ln(p(\mathbf{x}|\theta))}{\partial \theta^2}]$$

$$= -\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial}{\partial \theta}[\frac{\partial \ln(p(\mathbf{x}|\theta))}{\partial \theta}]] = -\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial}{\partial \theta}[\frac{\partial \eta^T}{\partial \theta}\frac{\partial \ln(p(\mathbf{x}|\eta))}{\partial \eta}]] = -\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial^2}{\partial \theta^2}[\frac{\partial \eta^T \partial \ln(p(\mathbf{x}|\eta))}{\partial \eta}]]$$

$$-\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial^2}{\partial \theta^2}\frac{\partial \eta^T \ln(p(\mathbf{x}|\eta))}{\partial \eta} + \frac{\partial \eta^T}{\partial \theta}\frac{\partial \eta}{\partial \theta}\frac{\partial^2 lnp(\mathbf{x}|\eta)}{\partial \eta^2}]$$

$$-\mathbb{E}_{p(\mathbf{x}|\theta)}[\frac{\partial \eta^T}{\partial \theta}\frac{\partial \eta}{\partial \theta}\frac{\partial^2 lnp(\mathbf{x}|\eta)}{\partial \eta^2}]$$

$$\frac{\partial \eta^T}{\partial \theta}\mathbf{F}(\eta)\frac{\partial \eta}{\partial \theta} = \mathbf{F}(\eta)^{-T}\mathbf{F}(\eta)\mathbf{F}(\eta)^{-1} = \mathbf{F}(\eta)$$

(d)

1 Practice

1. We see that the student's t-distribution looks like the typical Gaussian distribution more and more as the degree of freedom rises. When there is a big degree of freedom, the student-t distribution is shorter and less noticeable than the Gaussian distribution. I have coded this problem with Python and it is attached to this homework file in P1.py.

2. For the first task, the Beta distribution gets more symmetrical and gets bell-shaped and its mode moves closer to the distribution's center. I have coded this problem with Python and it is attached to this homework file in P2.py. For the second task, we can see the peak when we have different a

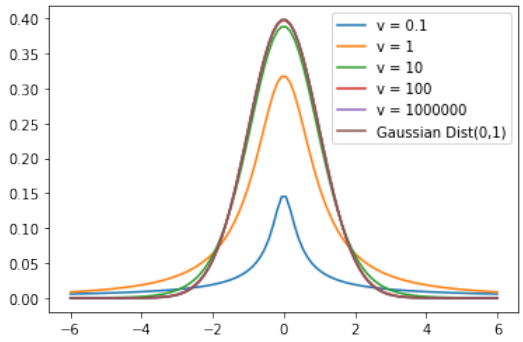

Figure 1: Practice 1: Value of Student-t Distribution with different degrees of freedom compared to Gaussian

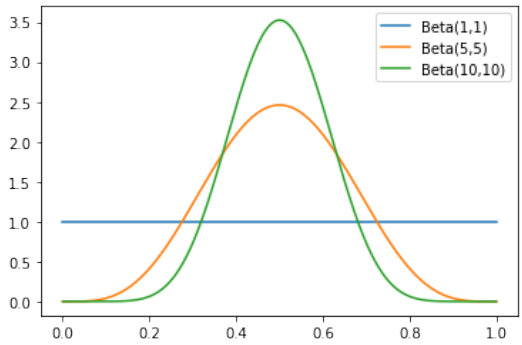

Figure 2: Practice 2: Density plots for Beta(1,1), Beta(5, 5) and Beta (10, 10)

and b at (2), and by increasing both a and b the mode moves forward to the same position as the last part and the peak increases and gets distance from the y=0 line.

3. Before injecting the noises, I ran the code multiple times and whenever the randomly generated samples were close together, these two density plots were completely aligned.
We observe that the estimated Gaussian density fits the data in its bell shape perfectly. However, the estimated Student-t resembles the Gaussian one very much but is less likely to lean towards the tails and where the sparse data are located, which shows it has a less tendency toward extreme values.
The graphic shows that, overall, the student-t distribution more closely fits the data than the Gaussian distribution. With the increase of the degrees of freedom, a Student-t distribution will have a distribution with heavier tails that mimics a normal distribution. As a result, the likelihood of the effects of extreme values and outliers decreases as the degrees of freedom increase. This makes the probability density more equally spread over the data set and less susceptible to outliers. With more freedom, the Student-t distribution can accommodate extreme values better.

After injecting the noises, we see that the Gaussian pdf no longer resembles the sample very well. As I said in the last part, Gaussian distribution shows some tendency towards extreme points, and we have injected three extreme points on the right side. We can see that still, the Student-t distribution remains virtually unchanged because it is more tolerant in accommodating and ignoring extreme values. After the data became more heavy-tailed, the Student-T distribution captures and provides a better fit to data better than the Gaussian distribution.

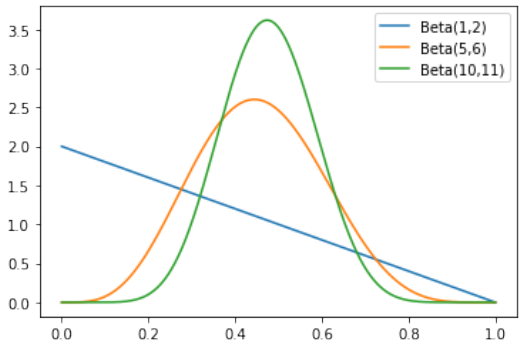

Figure 3: Practice 2: Density plots for Beta(1, 2), Beta(5,6) and Beta(10, 11)

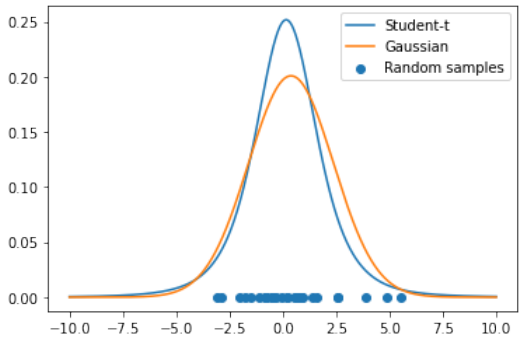

Figure 4: Practice 3: Density plots for Gaussian and student-t of the randomly drawn samples (shown in scatters)

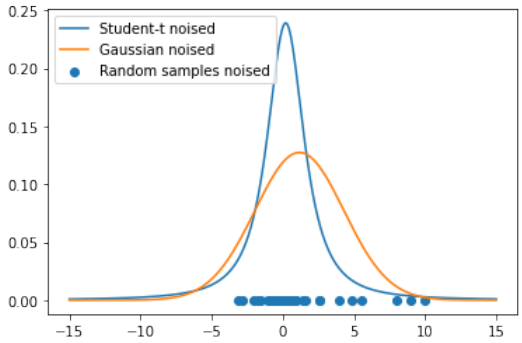

Figure 5: Practice 3: Density plots for Gaussian and student-t of the randomly drawn samples (shown in scatters) with added noise