Jared Yu
Midterm
Data Mining

References: [1], [2]

1. The Kullback-Leibler divergence of g from f is

$$D(f,g) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \log\left[\frac{f(\mathbf{x})}{g(\mathbf{x})}\right] d\mathbf{x}.$$

Here, $g(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, therefore $D(f,g)$ can be written as,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \log\left[\frac{f(\mathbf{x})}{(2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}\right] d\mathbf{x},$$

where $\mathbf{x}$ has is a $k \times 1$ vector. Expanding this leads to the following:

$$\cdots = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \left\{\log f(\mathbf{x}) - \log\left\{(2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]\right\}\right\} d\mathbf{x}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \left\{\log f(\mathbf{x}) + \frac{k}{2}\log 2\pi + \frac{1}{2}\log\det\boldsymbol{\Sigma} + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} d\mathbf{x}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \log f(\mathbf{x}) + f(\mathbf{x})\frac{k}{2}\log 2\pi + f(\mathbf{x})\frac{1}{2}\log\det\boldsymbol{\Sigma} + f(\mathbf{x})\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x}$$

The goal is to minimize this function; therefore, we need to take the derivative w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To simplify the function, $f(\mathbf{x}) \log f(\mathbf{x}) + f(\mathbf{x})\frac{k}{2}\log 2\pi$ will be left out of it for the remainder of the steps, since it will zero out anyways due to the derivative.

$$\frac{\partial D(f,g)}{\partial \boldsymbol{\mu}} = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \log\det\boldsymbol{\Sigma} + f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x}$$

In this derivative, the $f(\mathbf{x}) \log\det\boldsymbol{\Sigma}$ will also zero out due to the lack of a $\boldsymbol{\mu}$ term, and so it will be left out before expanding $(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$.

$$= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x}$$

$$= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) d\mathbf{x}$$

$$= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x})[\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}] d\mathbf{x}$$

$$= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x})[\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}] d\mathbf{x}$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \left[\frac{\partial}{\partial \boldsymbol{\mu}}(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}) - 2\frac{\partial}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}) + \frac{\partial}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right] d\mathbf{x}$$

In the above equation, $\frac{\partial}{\partial \mu}(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}) = 0$ and $\frac{\partial}{\partial \mu}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}) = \boldsymbol{\Sigma}^{-1}\mathbf{x}$, but $\frac{\partial}{\partial \mu}(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$ is more complicated. It will rely on the following,

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$

Therefore,

$$\frac{\partial}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) = (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^{\top})\boldsymbol{\mu} = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},$$

since the covariance matrix from the normal distribution is symmetric. Continuing from before:

$$\begin{aligned}
\cdots &= \frac{1}{2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})[-2\boldsymbol{\Sigma}^{-1}\mathbf{x} + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]d\mathbf{x} \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - f(\mathbf{x})\boldsymbol{\Sigma}^{-1}\mathbf{x}d\mathbf{x} \\
&= \boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{\mu}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})d\mathbf{x} - \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\mathbf{x}d\mathbf{x}\right\} \\
&= \boldsymbol{\Sigma}^{-1}\{\boldsymbol{\mu} - E_{f(\mathbf{x})}[\mathbf{x}]\} \overset{\text{set to}}{=} 0 \\
&\rightarrow \boldsymbol{\Sigma}^{-1}\{\boldsymbol{\mu} - E_{f(\mathbf{x})}[\mathbf{x}]\} = 0 \\
&\rightarrow \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\{\boldsymbol{\mu} - E_{f(\mathbf{x})}[\mathbf{x}]\} = \boldsymbol{\Sigma}0 \\
&\boxed{\rightarrow E_{f(\mathbf{x})}[\mathbf{x}] = \boldsymbol{\mu}}
\end{aligned}$$

Next, we need to find the derivative of $D(f, g)$ w.r.t. $\boldsymbol{\Sigma}$.

$$\begin{aligned}
\frac{\partial D(f,g)}{\partial \boldsymbol{\Sigma}} &= \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\log\det\boldsymbol{\Sigma} + f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} \\
&= \frac{1}{2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\frac{1}{\det\boldsymbol{\Sigma}}\frac{\partial}{\partial \boldsymbol{\Sigma}}\det\boldsymbol{\Sigma} + f(\mathbf{x})\frac{\partial}{\partial \boldsymbol{\Sigma}}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x}
\end{aligned}$$

To find $\frac{\partial}{\partial \boldsymbol{\Sigma}}\det\boldsymbol{\Sigma}$ and $\frac{\partial}{\partial \boldsymbol{\Sigma}}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, the following identities will be used,

$$\frac{\partial\det\mathbf{X}}{\partial\mathbf{X}} = \det(\mathbf{X})(\mathbf{X}^{-1})^{\top} \text{ and } \frac{\partial\mathbf{a}^{\top}\mathbf{X}^{-1}\mathbf{b}}{\partial\mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1})^{\top} = -(\mathbf{X}^{-1})^{\top}\mathbf{a}\mathbf{b}^{\top}(\mathbf{X}^{-1})^{\top}.$$

So, continuing from before leads to the following:

$$\begin{aligned}
\cdots &= \frac{1}{2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\frac{1}{\det\boldsymbol{\Sigma}}\det(\boldsymbol{\Sigma})(\boldsymbol{\Sigma}^{-1})^{\top} - f(\mathbf{x})(\boldsymbol{\Sigma}^{-1})^{\top}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}(\boldsymbol{\Sigma}^{-1})^{\top}d\mathbf{x} \\
&= \frac{1}{2}\left\{\boldsymbol{\Sigma}^{-1} - \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}d\mathbf{x}\right\} \overset{\text{set to}}{=} 0 \\
&\rightarrow \boldsymbol{\Sigma}^{-1} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}d\mathbf{x} \\
&\rightarrow \boldsymbol{\Sigma} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}d\mathbf{x} \\
&\boxed{\rightarrow \boldsymbol{\Sigma} = E_{f(\mathbf{x})}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}]}
\end{aligned}$$

To confirm that these are truly minimizing $D(f, g)$, the second derivatives are also required. This is difficult in the case of $\boldsymbol{\Sigma}$, since it would require that a tensor be shown to as positive definite. Through an email conversation with the professor, it was mentioned that such a step is possibly not required. In the case of $\boldsymbol{\mu}$, however, it is much simpler.

$$\frac{\partial^2 D(f,g)}{\partial\boldsymbol{\mu}^{\top}\boldsymbol{\partial\mu}} = \frac{\partial}{\partial\boldsymbol{\mu}^{\top}}\boldsymbol{\Sigma}^{-1}\{\boldsymbol{\mu} - E_{f(\mathbf{x})}[\mathbf{x}]\} = \frac{\partial}{\partial\boldsymbol{\mu}^{\top}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

$$= \frac{\partial}{\partial \boldsymbol{\mu}^\top} \begin{bmatrix} \sum_{i=1}^{d} a_{1i}\mu_i \\ \sum_{i=1}^{d} a_{2i}\mu_i \\ \vdots \\ \sum_{i=1}^{d} a_{di}\mu_i \end{bmatrix}$$

In the above equation, a_{ji} corresponds to the j^{th} row and i^{th} column of $\mathbf{\Sigma}^{-1}$.

$$\cdots = \begin{bmatrix} \frac{\partial}{\partial \mu_1}\left(\sum_{i=1}^{d} a_{1i}\mu_i\right) & \cdots & \frac{\partial}{\partial \mu_1}\left(\sum_{i=1}^{d} a_{di}\mu_i\right) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \mu_d}\left(\sum_{i=1}^{d} a_{1i}\mu_i\right) & \cdots & \frac{\partial}{\partial \mu_d}\left(\sum_{i=1}^{d} a_{di}\mu_i\right) \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{d1} \\ \vdots & \ddots & \vdots \\ a_{1d} & \cdots & a_{d} \end{bmatrix} = \mathbf{\Sigma}^{-1}.$$

The $\mathbf{\Sigma}^{-1}$ is a positive semi-definite matrix and so that shows that $E_{\boldsymbol{f}(\mathbf{x})}[\mathbf{x}] = \boldsymbol{\mu}$ is likely a minimum.

References: [3]. [4]

2. The conditional probability for the n-dimensional multivariate Bernoulli is given as follows,

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

and let $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ be a set of k samples independently drawn from $P(\mathbf{x}|\boldsymbol{\theta})$.

a. Let $\boldsymbol{s} = (s_1, \cdots, s_n)^\top$ be the sum of the set of k samples. Here, $P(\mathbf{x}|\boldsymbol{\theta})$ can be rewritten to include an index j representing the j^{th} sample from $\mathcal{D}$, where $1 \le j \le k$.

$$P\left(\mathbf{x}_j|\boldsymbol{\theta}\right) = \prod_{i=1}^{n} \theta_i^{x_{ij}}(1-\theta_i)^{1-x_{ij}}$$

So, $\boldsymbol{s}$ can be written as the following,

$$\boldsymbol{s} = (s_1, \cdots, s_n)^\top = \begin{pmatrix} \sum_{j=1}^{k} x_{1j} & \sum_{j=1}^{k} x_{2j} & \cdots & \sum_{j=1}^{k} x_{nj} \end{pmatrix}^\top.$$

Here, $P(\mathcal{D}|\boldsymbol{\theta})$, can be expressed as follows:

$$P(\mathcal{D}|\boldsymbol{\theta}) = P(\mathbf{x}_1, \cdots, \mathbf{x}_k|\boldsymbol{\theta}) = \prod_{j=1}^{k}\prod_{i=1}^{n}\left[\theta_i^{x_{ij}}(1-\theta_i)^{1-x_{ij}}\right]$$

$$= \left\{\prod_{i=1}^{n}\left[\theta_i^{x_{i1}}(1-\theta_i)^{1-x_{i1}}\right]\right\} \times \left\{\prod_{i=1}^{n}\left[\theta_i^{x_{i2}}(1-\theta_i)^{1-x_{i2}}\right]\right\} \times \cdots \times \left\{\prod_{i=1}^{n}\left[\theta_i^{x_{ik}}(1-\theta_i)^{1-x_{ik}}\right]\right\}$$

$$= \prod_{i=1}^{n} \theta_i^{\sum_{j=1}^{k} x_{ij}} (1-\theta_i)^{k-\sum_{j=1}^{k} x_{ij}} \boxed{= \prod_{i=1}^{n} \theta_i^{s_i} (1-\theta_i)^{k-s_i}} \ \blacksquare$$

b. The Bayes formula can be expressed as follows,

$$posterior = \frac{likelihood \times prior}{evidence}.$$

Adapting this formula to the current problem leads to the following,

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})}.$$

To find the posterior, it requires the likelihood, prior, and evidence. The likelihood has been shown in part a). So, for this problem the prior and evidence need to be found before the posterior can be shown. It is given that the distribution of the prior is uniform. Therefore, each of the $\theta_i \sim Uniform(0,1)$ for $i = 1, \cdots, n$. Then, $P(\boldsymbol{\theta})$ is the joint $P(\theta_1, \cdots, \theta_n)$, which equates to 1 since the uniform probability density function (pdf) is simply 1 in this case. It has been explained that the prior, $P(\boldsymbol{\theta}) = 1$, so next the evidence $P(\mathcal{D})$ must be found. The evidence can be written as the marginal probability of the joint probability of the likelihood which can be seen as follows:

$$P(\mathcal{D}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(\mathcal{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(\mathcal{D}|\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$= \int_{0}^{1} \cdots \int_{0}^{1} \prod_{i=1}^{n} \theta_i^{s_i} (1-\theta_i)^{k-s_i} \, d\theta_1 \cdots d\theta_n = \prod_{i=1}^{n} \int_{0}^{1} \theta_i^{s_i} (1-\theta_i)^{k-s_i} \, d\theta_i$$

From here, we can look at the pdf of the Beta distribution which looks quite similar:

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \, dx = 1$$

$$\rightarrow \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Then by equating this version of the Beta distribution, we can find equivalent values for α and β from $\prod_{i=1}^{n} \int_0^1 \theta_i^{s_i} (1-\theta_i)^{k-s_i} \, d\theta_i$.

$$\alpha - 1 = s_i \rightarrow \alpha = s_i + 1$$

$$\beta - 1 = k - s_i \rightarrow \beta = k - s_i + 1$$

Using the above steps, we can transform our current $P(\mathcal{D})$ into something more manageable:

$$\cdots = \prod_{i=1}^{n} \frac{\Gamma(s_i+1)\Gamma(k-s_i+1)}{\Gamma(k+2)} = \prod_{i=1}^{n} \frac{s_i!\,(k-s_i)!}{(k+1)!}$$

Now that $P(\mathcal{D})$ has been found, we can go back to the posterior:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})} = \frac{\prod_{i=1}^{n} \theta_i^{s_i}(1-\theta_i)^{k-s_i} \times (1)}{\prod_{i=1}^{n} \frac{s_i!\,(k-s_i)!}{(k+1)!}} = \boxed{\prod_{i=1}^{n} \frac{(k+1)!}{s_i!\,(k-s_i)!} \theta_i^{s_i}(1-\theta_i)^{k-s_i}} \ \blacksquare$$

c. Letting $n = 1$ and $k = 1$ $P(\boldsymbol{\theta}|\mathcal{D})$ becomes the following:

$$\frac{2!}{s!\,(1-s)!} \theta^s (1-\theta)^{1-s}$$

In such a situation, there are two possibilities for s,

$$\begin{cases} s = 0: & 2(1-\theta) \\ s = 1: & 2\theta, \end{cases}$$

where $0 \leq \theta \leq 1$. The plot of the density can be seen below in Figure 1. The case for $s = 0$ is in red, and $s = 1$ is in blue. The plot shows $P(\theta|\mathcal{D})$ on the y-axis and θ on the x-axis.

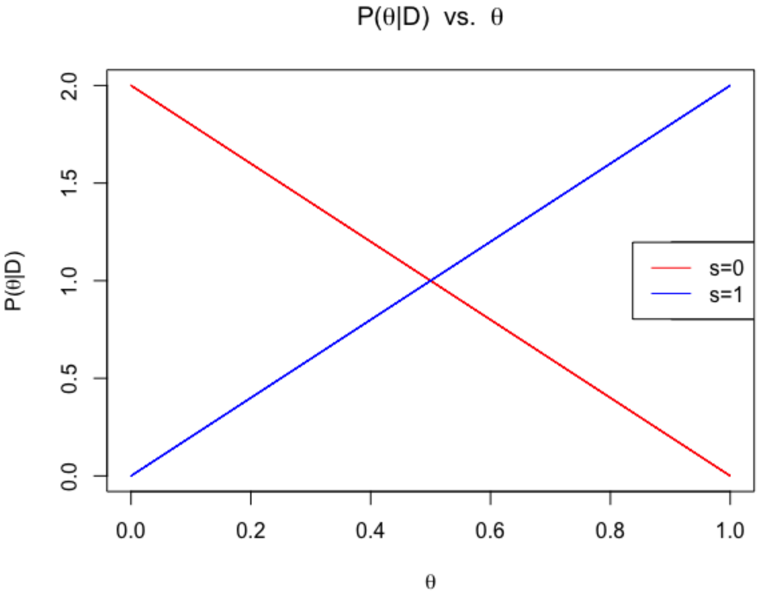

Figure 1 The above figure shows $P(\theta|\mathcal{D})$ vs. θ in the two cases where $s = 0$ (red) and $s = 1$ (blue).

d. The question is asking to find $P(\mathbf{x}|\mathcal{D})$ by integrating $P(\mathbf{x}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{D})$ over $\boldsymbol{\theta}$.

$$\int_0^1 \cdots \int_0^1 P(\mathbf{x}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta} = \int_0^1 \cdots \int_0^1 \prod_{i=1}^n \theta_i^{x_i}(1-\theta_i)^{1-x_i} \prod_{i=1}^n \frac{(k+1)!}{s_i!\,(k-s_i)!} \theta_i^{s_i}(1-\theta_i)^{k-s_i}\, d\boldsymbol{\theta}$$

$$= \int_0^1 \cdots \int_0^1 \prod_{i=1}^n \theta_i^{x_i}(1-\theta_i)^{1-x_i} \frac{(k+1)!}{s_i!\,(k-s_i)!} \theta_i^{s_i}(1-\theta_i)^{k-s_i}\, d\boldsymbol{\theta}$$

$$= \prod_{i=1}^n \int_0^1 \theta_i^{x_i}(1-\theta_i)^{1-x_i} \frac{(k+1)!}{s_i!\,(k-s_i)!} \theta_i^{s_i}(1-\theta_i)^{k-s_i}\, d\theta_i$$

$$= \prod_{i=1}^n \frac{(k+1)!}{s_i!\,(k-s_i)!} \int_0^1 \theta_i^{x_i}(1-\theta_i)^{1-x_i}\theta_i^{s_i}(1-\theta_i)^{k-s_i}\, d\theta_i$$

$$= \prod_{i=1}^n \frac{(k+1)!}{s_i!\,(k-s_i)!} \int_0^1 \theta_i^{x_i+s_i}(1-\theta_i)^{1-x_i+k-s_i}\, d\theta_i$$

Again, we can use the Beta distribution like in part b):

$$\rightarrow \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Then by equating this version of the Beta distribution, we can find equivalent values for α and β from $\int_0^1 \theta_i^{x_i+s_i}(1-\theta_i)^{1-x_i+k-s_i}\, d\theta_i$:

$$\alpha - 1 = x_i + s_i \rightarrow \alpha = x_i + s_i + 1$$
$$\beta - 1 = 1 - x_i + k - s_i \rightarrow \beta = 2 - x_i + k - s_i$$

This gives us the following:

$$\begin{aligned}\cdots &= \prod_{i=1}^{n} \frac{(k+1)!}{s_i!\,(k-s_i)!} \times \frac{\Gamma(x_i+s_i+1)\Gamma(2-x_i+k-s_i)}{\Gamma(k+3)} \\ &= \prod_{i=1}^{n} \frac{(k+1)!}{s_i!\,(k-s_i)!} \times \frac{(x_i+s_i)!\,(1-x_i+k-s_i)!}{(k+2)!} \\ &= \prod_{i=1}^{n} \frac{1}{(k+2)} \frac{(x_i+s_i)!}{s_i!} \frac{(1-x_i+k-s_i)!}{(k-s_i)!}\end{aligned}$$

Looking at the above equation, it can be simplified in the following cases for x_i:

$$\frac{(x_i+s_i)!}{s_i!} = (s_i+1)^{x_i} = \begin{cases} 1 & \text{if } x_i = 0 \\ s_i + 1 & \text{if } x_i = 1 \end{cases}$$

$$\frac{(1-x_i+k-s_i)!}{(k-s_i)!} = (k-s_i+1)^{1-x_i} = \begin{cases} k - s_i + 1 & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \end{cases}$$

Plugging these functions back into the pervious equation:

$$\cdots = \prod_{i=1}^{n} \frac{1}{(k+2)} (s_i+1)^{x_i} (k-s_i+1)^{1-x_i}$$

$$= \prod_{i=1}^{n} \frac{(s_i+1)^{x_i}(k-s_i+1)^{1-x_i}}{(k+2)^{x_i}(k+2)^{1-x_i}} = \prod_{i=1}^{n} \left(\frac{s_i+1}{k+2}\right)^{x_i} \left(\frac{k-s_i+1}{k+2}\right)^{1-x_i}$$

$$\boxed{= \prod_{i=1}^{n} \left(\frac{s_i+1}{k+2}\right)^{x_i} \left(1 - \frac{s_i+1}{k+2}\right)^{1-x_i}} \;\blacksquare$$

e. The formulas for $P(\mathbf{x}|\mathcal{D})$ and $P(\mathbf{x}|\boldsymbol{\theta})$ are as follows,

$$P(\mathbf{x}|\mathcal{D}) = \prod_{i=1}^{n} \left(\frac{s_i+1}{k+2}\right)^{x_i} \left(1 - \frac{s_i+1}{k+2}\right)^{1-x_i} \text{ and } P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

It can be seen that between the two formulas, the only difference is that $\left(\frac{s_i+1}{k+2}\right)$ and θ_i are interchanged. In $P(\mathbf{x}|\boldsymbol{\theta})$, θ_i is the probability that $x_i = 1$, so it must be between 0 and 1. Looking to $P(\mathbf{x}|\mathcal{D})$, $\left(\frac{s_i+1}{k+2}\right)$ must also be a fraction (similar to a probability) between 0 and 1, since s_i is at most k, and the value can't go lower than 0 or exceed 1. In $P(\mathbf{x}|\mathcal{D})$, we can think of $\mathbf{x}$ as a new sample and so it is the probability distribution o $\mathbf{x}$, given the old evidence $\mathcal{D}$. It would be calculated after the posterior has been found and new evidence is discovered. On the other hand, $P(\mathbf{x}|\boldsymbol{\theta})$ can be thought of as the probability of a sample $\mathbf{x}$, given the prior information $\boldsymbol{\theta}$.

To find $\widehat{\boldsymbol{\theta}}$, the effective Bayesian estimate of $\boldsymbol{\theta}$ is the expected value of the $\boldsymbol{\theta}|\mathcal{D}$ distribution or posterior distribution, where $\theta_i|\mathcal{D} \sim Beta(s_i + 1, k - s_i + 1)$, since it minimizes the quadratic loss. We know that $\theta_i|\mathcal{D}$ follows a Beta distribution, since in part b) it shows that the posterior is a joint Beta distribution where $\alpha = s_i + 1$ and $\beta = k - s_i + 1$.

So the expected value of the $\boldsymbol{\theta}|\mathcal{D}$ distribution is $E(\theta_i|\mathcal{D}) = \frac{s_i+1}{(s_i+1)+(k-s_i+1)} = \frac{s_i+1}{k+2}$. Therefore, $\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \frac{s_1+1}{k+2} & \frac{s_2+1}{k+2} & \cdots & \frac{s_n+1}{k+2} \end{pmatrix}^{\top}$.

3. The problem asks us to analyze a dataset `midterm_exam_data.txt`.
   a. For this problem, the number of different fits tried were for $n \in \{1, \cdots, 10\}$. This leads to 10 different datasets and 10 different corresponding models based on the cosine basis function. The idea is to generate a design matrix as follows,

$$\mathbf{X} = \begin{bmatrix} \frac{1}{2} & \cos x_1 & \cos 2x_1 & \cdots & \cos nx_1 \\ \frac{1}{2} & \cos x_2 & \cos 2x_x & \cdots & \cos nx_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & \cos x_m & \cos 2x_m & \cdots & \cos nx_m \end{bmatrix}.$$

The above design matrix shows the possibility of adding up to n cosine basis functions to as a potential design matrix, where in this case $m = 60$ corresponds to the total number of samples in the dataset. In this case, the number used is $n = 10$ for the 10 different models. Each of the 10 models require their own matching dataset.

The problem then becomes a matter of estimating $a_0, \cdots, a_n$, similar to estimating the β_j terms in a multiple linear regression problem. Therefore, a similar method was used where the $\hat{\alpha}_j$ terms corresponded with the least squares normal equations (i.e., $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$) used in multiple linear regression.

After fitting the 10 models to the data, their corresponding R^2_{Adj} was calculated. The R^2_{Adj} has the following formula,

$$R^2_{Adj} = 1 - \frac{\frac{SS_{res}}{df_e}}{\frac{SS_{tot}}{df_t}},$$

where $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$, $SS_{tot} = \sum_i (y_i - \bar{y})^2$, $\frac{1}{n}\sum_i y_i$, $df_e = m - p - 1$, $df_t = m - 1$ (where here m refers to the number of observations in the dataset i.e. 50), and p is the number of parameters excluding the constant term (e.g., for $n = 3, p = 3$, etc.).

This can be seen below in Figure 2. In the figure,

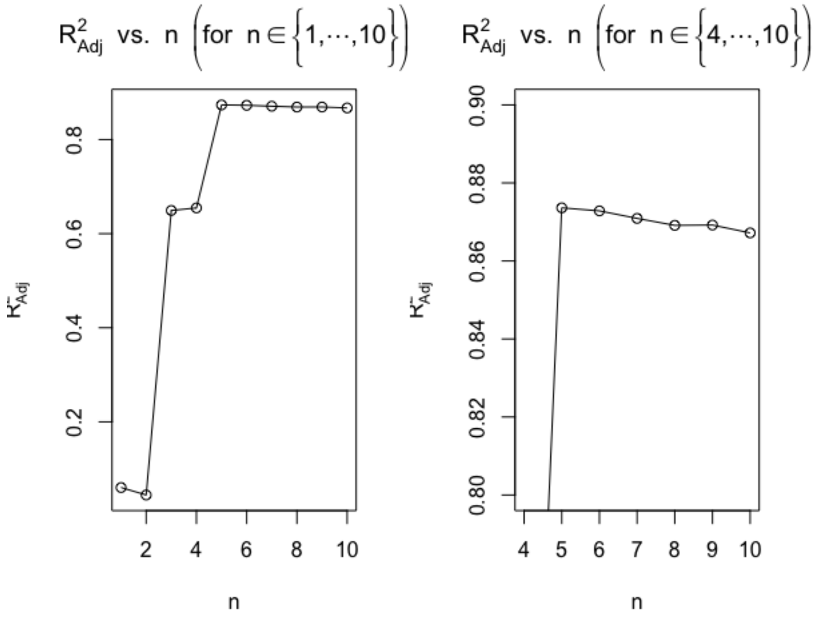

Figure 2 The above figure shows the R^2_{Adj} vs. n for the ten models. The left plot shows all ten models, while the right plot zooms in on the last 6 models to help show that the 5th model has the highest peak.

Based on Figure 2, it is apparent that the best model is $n = 5$, based on the R^2_{Adj} metric. It not only has the highest R^2_{Adj} compared to the other models, but from $n = 1, \cdots, 4$, the R^2_{Adj} is considerably lower. Then, from $n = 6, \cdots, 10$, the R^2_{Adj} is strictly lower. This makes it certain that from amongst these models $n = 5$ is the best choice based on R^2_{Adj}. Adding more complexity to the model does not improve the performance when using R^2_{Adj}. Therefore, the most likely value of n is $\boxed{n = 5}$.

This final model can be seen plotted over the data below in Figure 3. The fit appears quite smooth and matches the dataset well.

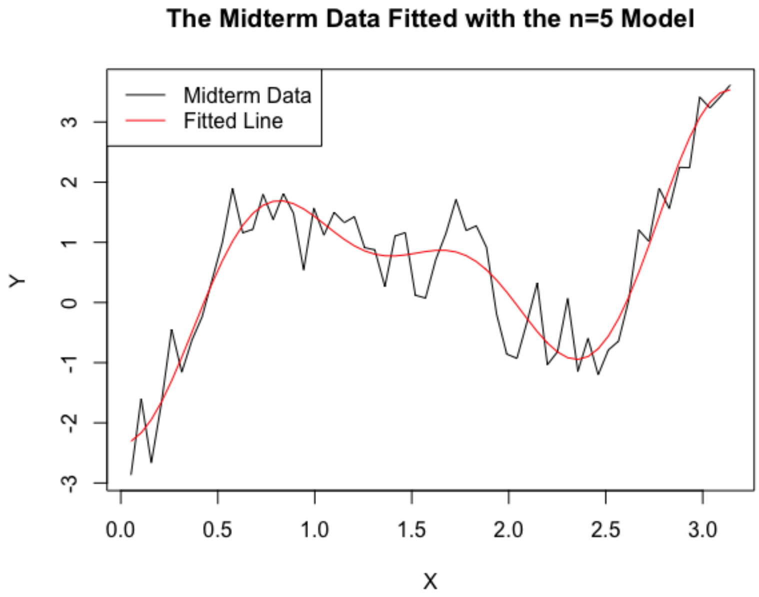

Figure 3 The above figure shows the sample data plotted on the x and y axes, with the chosen fitted model for $n = 5$ *in red.*

b. It is stated that the sum of the cosines has some added Gaussian noise. That is, the data is a linear function of the cosine basis functions in addition to a random error. This random error, ε is normally distributed with mean zero and some unknown variance σ^2, i.e., $\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. This random error can be understood from our sample data through the residual terms, $e_i = y_i - \hat{y}_i$. Taking the sample mean of all the e_i terms, we get $-7.760267 \times 10^{-17}$ which is approximately 0. To estimate the level of noise, we can estimate σ^2 with $\boxed{\hat{\sigma}^2 = \frac{SS_{Res}}{m-p} \approx 0.2482}$.

References

[1] https://en.wikipedia.org/wiki/Multivariate_normal_distribution

[2] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

[3] Duda, R. O. (2000). R. O. Duda's P. E. Hart's D. G. Stork's Pattern Classification (Pattern Classification (2nd Edition) [Hardcover])(2000) (2 edition). Wiley-Interscience.

[4] https://en.wikipedia.org/wiki/Beta_distribution

Code

```r
### Problem 2 (c)
# plot density vs theta
thetas <- seq(0, 1, length.out = 1e4)
s0 <- function(theta) { return(2 * (1 - theta)) }
s1 <- function(theta) { return(2 * theta) }
plot(thetas, s0(thetas), type = 'l', col = 'red',
     main = latex2exp::TeX('$P(\\theta | D) \\;vs.\\;\\theta$'),
     xlab = latex2exp::TeX('$\\theta$'),
     ylab = latex2exp::TeX('$P(\\theta | D)$'))
lines(thetas, s1(thetas), col = 'blue')
legend("right",
       legend = c(latex2exp::TeX('$s=0$'),
                  latex2exp::TeX('$s=1$')),
       col = c('red', 'blue'), lty = c(1,1))

# Load data
polynomial_data <- read.csv('xid-94690945_2.txt', header = FALSE, sep = "")
colnames(polynomial_data) <- c("X", "Y")

# Create initial variables
m <- nrow(polynomial_data)
X <- polynomial_data[,1]
Y <- polynomial_data[,2]

# Calculate the nth degree for X
nth_degree <- function(x, n) {
  return(cos(n * x))
}

nth <- 10
polynomial_list <- lapply(1:nth, function(y) nth_degree(x=X, n=y))

# Create the matrices for the nth degree polynomials
halves <- rep(0.5, nth)
# data_matrix = polynomial_list[1:3]
create_design_matrix <- function(halves_vector=halves, data_matrix) {
  design_mat <- cbind(halves_vector, do.call(cbind, data_matrix))
  return(design_mat)
}
# polynomial_list[1:2] # list these out in 1:1, 1:2, 1:3, etc.
polynomial_degree_vec <- sapply(seq(1, nth), function(x) c(1,x))
polynomial_degree_groups <- lapply(1:nth, function(x) polynomial_list[1:x])

# List of the different design matrices for various degree polynomials
polynomial_design_matrix_list <- lapply(polynomial_degree_groups,
  function(x) create_design_matrix(data_matrix = x))

# Calculate the beta hat
beta_hat <- function(X, y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}

beta_hat_list <- lapply(polynomial_design_matrix_list,
  function(x) beta_hat(X = x, y = Y))

# Calculate the adjusted R-squared
adj_r_squared <- function(X, betahat, y) {
  y_bar <- mean(y)
  y_hat <- X %*% betahat

  SS_tot <- sum((y - y_bar)^2)
  SS_res <- sum((y - y_hat)^2)

  n <- length(y)
  p <- ncol(X) - 1
  df_e <- n - p - 1
  df_t <- n - 1

  adj_r_square <- 1 - ((SS_res / df_e) / (SS_tot / df_t))
  return(adj_r_square)
}

adj_r_squared_list <- mapply(function(a, b) adj_r_squared(X = a, betahat = b, y = Y),
  polynomial_design_matrix_list, beta_hat_list)

par(mfrow = c(1,2))
plot(1:nth, adj_r_squared_list, type='l',
     main = latex2exp::TeX('$R_{Adj}^2 \\; vs. \\; n \\; \\left(for \\;n \\in \\left{1,\\cdots,10 \\right}
\\right)$'),
     xlab = latex2exp::TeX('$n$'), ylab = latex2exp::TeX('$R_{Adj}^2$'))
points(1:nth, adj_r_squared_list)

plot(4:nth, adj_r_squared_list[4:nth], type='l', ylim = c(0.8, 0.9),
     main = latex2exp::TeX('$R_{Adj}^2 \\; vs. \\; n \\; \\left(for \\;n \\in \\left{4,\\cdots,10 \\right}
\\right)$'),
     xlab = latex2exp::TeX('$n$'), ylab = latex2exp::TeX('$R_{Adj}^2$'))
points(4:nth, adj_r_squared_list[4:nth])

### Chosen model: n=5
y_hat <- polynomial_design_matrix_list[[5]] %*% beta_hat_list[[5]]
plot(X,Y, type = 'l',
     main = 'The Midterm Data Fitted with the n=5 Model')
legend("topleft", legend = c('Midterm Data', 'Fitted Line'),
       col = c('black', 'red'), lty = c(1,1))
lines(X,y_hat, col = 'red')

e <- Y - y_hat
mean(e) # correct
p <- length(beta_hat_list[[5]])
SSE <- sum(e^2)
sigma_hat_2 <- SSE / (m - p)
```