Machine Learning 2 **Due: April 05, 2016**

Homework 1

Instructor: Ke Tran Email: m.k.tran@uva.nl

Student: Andrea Jemmett UvA ID: 11162929
Collaborators: N/A Email: andreajemmett@gmail.com

Problem 1. Consider two random vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^n$ having Gaussian distribution $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z})$. Consider random vector $\mathbf{y} = \mathbf{x} + \mathbf{z}$. Derive mean and covariance of $p(\mathbf{y})$.

Problem 2. Given a set of N observations $\mathcal{X} = \{x_1, \ldots, x_N\}$. Assume that $x_i \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

1. Write down the likelihood of the data $p(\mathcal{X}|\mu, \sigma^2)$;

$$p(\mathcal{X}|\mu, \sigma^2) = \prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \tag{1}$$

2. Write down the posterior $p(\mu|\mathcal{X}, \sigma^2, \mu_0, \sigma_0^2)$;

$$\begin{aligned} p(\mu|\mathcal{X}, \sigma^2, \mu_0, \sigma_0^2) &= p(\mathcal{X}|\mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \\ &= \prod_{i=1}^{N} p(x_i|\mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \end{aligned} \tag{2}$$

3. Show that $p(\mu|\mathcal{X}, \sigma^2, \mu_0, \sigma_0^2)$ is a Gaussian distribution $\mathcal{N}(\mu|\mu_N, \sigma_N^2)$ and find the values of μ_N and σ_N^2;

$$
\begin{aligned}
p(\mu|\mathcal{X},\sigma^2,\mu_0,\sigma_0^2) &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\\
&= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\\
&= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}x_i^2-\sum_{i=1}^{N}2x_i\mu+N\mu^2\right)-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0+\mu_0^2)\right\}\\
&= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\exp\left\{-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)+\mu\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i+\frac{\mu_0}{\sigma_0^2}\right)+\text{const}\right\}\\
&= \frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\exp\left\{\underbrace{-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)}_{-\frac{1}{2\sigma_N^2}}+\mu\underbrace{\left(\frac{N}{\sigma^2}\mu_{\mathrm{ML}}+\frac{\mu_0}{\sigma_0^2}\right)}_{\frac{1}{\sigma_N^2}\mu_N}+\text{const}\right\}
\end{aligned}
\tag{3}
$$

where $\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{i=1}^{N}x_i$ is the sample mean, *const* are terms not dependent on μ and it is a Gaussian distribution (because of the form of coefficients entering quadratic μ^2 and linear μ terms) with mean and variance given by

$$
\begin{aligned}
\frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\\
\mu_N &= \left(\frac{N\mu_{\mathrm{ML}}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\sigma_N^2\\
&= \left(\frac{N\mu_{\mathrm{ML}}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\frac{\sigma_0^2\sigma^2}{\sigma^2+N\sigma_0^2}\\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}
\end{aligned}
\tag{4}
$$

4. Derive the maximum a posterior solution for μ;

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\log p(\mu|\mathcal{X},\sigma^2,\mu_0,\sigma_0^2) &= \frac{\partial}{\partial\mu}\log\left(\frac{1}{2\pi\sqrt{\sigma^2\sigma_0^2}}\right)-\frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]-\frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]\\
&= \frac{N}{2\sigma^2}\frac{\partial}{\partial\mu}\mu^2+\frac{1}{2\sigma_0^2}\frac{\partial}{\partial\mu}\mu^2-\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i\frac{\partial}{\partial\mu}\mu-\frac{\mu_0}{\sigma_0^2}\frac{\partial}{\partial\mu}\mu+\frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N}x_i^2+\frac{1}{2\sigma_0^2}\mu_0^2\right]\\
&= \frac{N}{\sigma^2}\mu+\frac{1}{\sigma_0^2}\mu-\frac{N\mu_{\mathrm{ML}}}{\sigma^2}-\frac{\mu_0}{\sigma_0^2}\\
&= \frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\mu-\frac{N\sigma_0^2\mu_{\mathrm{ML}}+\mu_0\sigma^2}{\sigma^2\sigma_0^2}=0
\end{aligned}
\tag{5}
$$

we can then solve for μ

$$\begin{aligned}
\hat{\mu}_{\mathrm{MAP}} &= \frac{N\sigma_0^2\mu_{\mathrm{ML}} + \mu_0\sigma^2}{\sigma^2\sigma_0^2}\frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2} \\
&= \frac{N\sigma_0^2\mu_{\mathrm{ML}} + \sigma^2\mu_0}{N\sigma_0^2+\sigma^2} \\
&= \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 \\
&= \mu_N
\end{aligned} \tag{6}$$

5. Derive expressions for sequential update of μ_N and σ_N^2;

   First define $\mu_N^{(N)}$ as the estimated μ_N using N samples. Then we can write:

$$\begin{aligned}
\mu_N^{(N)} &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}} \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\frac{1}{N}\sum_{i=1}^{N} x_i \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\sum_{i=1}^{N} x_i \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(x_N + \sum_{i=1}^{N-1} x_i\right) \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\left[x_N + (N-1)\mu_{\mathrm{ML}}^{(N-1)}\right] \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{(N-1)\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}^{(N-1)} + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N \\
&= \frac{1}{N\sigma_0^2+\sigma^2}(\sigma^2\mu_0 + (N-1)\sigma_0^2\mu_{\mathrm{ML}}^{(N-1)}) + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N \\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\left(\frac{\sigma^2}{(N-1)\sigma_0^2+\sigma^2}\mu_0 + \frac{(N-1)\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}^{(N-1)}\right) + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N \\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\mu_N^{(N-1)} + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N
\end{aligned} \tag{7}$$

6. Derive the same results (as in 5) starting from the posterior distribution $p(\mu|x_1,\ldots,x_{N-1})$, and multiplying by the likelihood function $p(x_N|\mu) = \mathcal{N}(x_N|\mu,\sigma^2)$.

Problem 3. Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ in which the covariance $\boldsymbol{\Sigma}$ is known and for which we wish to infer the mean $\boldsymbol{\mu}$ from a set of observations $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$.

1. Write down the likelihood of the data $p(\mathcal{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$;

$$
\begin{aligned}
p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{D} p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \prod_{i=1}^{D} \exp\left\{\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right\}
\end{aligned} \tag{8}
$$

2. Given a prior distribution $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, find the corresponding posterior distribution $p(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{X}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0) &= p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \Sigma_0) p(\mathcal{X}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_0|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right\} \\
&\quad \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right\} \\
&= (2\pi)^{-D} |\Sigma\Sigma_0|^{-1} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \sum_{i=1}^{D} \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}) \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right\}
\end{aligned} \tag{9}
$$

We then observe that for a symmetric matrix $\mathbf{A}$, holds that

$$
\mathbf{a}^T \mathbf{A} \mathbf{b} = \mathbf{b}^T \mathbf{A} \mathbf{a} \tag{10}
$$

and so

$$
\begin{aligned}
(\mathbf{a} - \mathbf{b})^T \mathbf{A} (\mathbf{a} - \mathbf{b}) &= (\mathbf{a}^T - \mathbf{b}^T) \mathbf{A} (\mathbf{a} - \mathbf{b}) \\
&= (\mathbf{a}^T - \mathbf{b}^T)(\mathbf{A}\mathbf{a} - \mathbf{A}\mathbf{b}) \\
&= \mathbf{a}^T \mathbf{A} \mathbf{a} - \mathbf{a}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{A} \mathbf{a} + \mathbf{b}^T \mathbf{A} \mathbf{b} \\
&= \mathbf{a}^T \mathbf{A} \mathbf{a} + \mathbf{b}^T \mathbf{A} \mathbf{b} - \mathbf{a}^T \mathbf{A} \mathbf{b} - \mathbf{a}^T \mathbf{A} \mathbf{b} \\
&= \mathbf{a}^T \mathbf{A} \mathbf{a} + \mathbf{b}^T \mathbf{A} \mathbf{b} - 2\mathbf{a}^T \mathbf{A} \mathbf{b}
\end{aligned} \tag{11}
$$

Because we know that both Σ and Σ_0 are symmetric, we can write the exponential term of the posterior as

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{D}\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i + \sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{D}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&= -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} - \frac{D}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i \\
&= -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} - \frac{D}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i \\
&= -\frac{1}{2}\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1} + D\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i\right) - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i
\end{aligned}
\tag{12}
$$

We can finally write the posterior

$$
\begin{aligned}
& p(\boldsymbol{\mu}|\mathcal{X},\Sigma,\boldsymbol{\mu}_0,\Sigma_0) = \\
& (2\pi)^{-D}|\Sigma\Sigma_0|^{-1}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1} + D\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i\right) - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\sum_{i=1}^{D}\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right\}
\end{aligned}
\tag{13}
$$

3. Show that the posterior $p(\boldsymbol{\mu}|\mathcal{X},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_N$ and covariance $\boldsymbol{\Sigma}_N$

$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{X},\Sigma,\boldsymbol{\mu}_0,\Sigma_0) &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1} + D\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i\right)\right\} && (14) \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N\right\} && (15) \\
&\propto \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N) && (16)
\end{aligned}
$$

4. Find $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$

$$
\begin{aligned}
\boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + D\boldsymbol{\Sigma}^{-1})^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i\right) \\
&= \left(\boldsymbol{\Sigma}_0^{-1} + D\boldsymbol{\Sigma}^{-1}\right)\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{D}\mathbf{x}_i\right)
\end{aligned}
\tag{17}
$$

Problem 4.

1. Show that the product of two Gaussians gives another (un-normalized) Gaussian

$$\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) = K^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c},\mathbf{C})$$

where $\mathbf{c} = \mathbf{C}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b})$ and $\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}$.

$$\begin{aligned}
\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) &= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\mathbf{A}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a})\right\} \\
&\quad \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\mathbf{B}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b})\right\} \\
&= (2\pi)^{-D}|\mathbf{AB}|^{-\frac{1}{2}}\exp\underbrace{\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) - \frac{1}{2}(\mathbf{x}-\mathbf{b})^\top\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b})\right\}}_{f(\mathbf{x})}
\end{aligned} \tag{18}$$

We can then develop the exponential term using the results of (11)

$$\begin{aligned}
f(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) - \frac{1}{2}(\mathbf{x}-\mathbf{b})^\top\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b}) \\
&= -\frac{1}{2}\mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{x}^\top\mathbf{B}^{-1}\mathbf{x} + \mathbf{x}^\top\mathbf{B}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b} \\
&= -\frac{1}{2}\mathbf{x}^\top(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x} + \mathbf{x}^\top(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b} \\
&= -\frac{1}{2}\mathbf{x}^\top\mathbf{C}^{-1}\mathbf{x} + \mathbf{x}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}
\end{aligned} \tag{19}$$

where the following substitution has been adopted

$$\begin{aligned}
\mathbf{C}^{-1} &= \mathbf{A}^{-1} + \mathbf{B}^{-1} \\
\mathbf{C} &= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} \\
\\
\mathbf{C}^{-1}\mathbf{c} &= \mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b} \\
\mathbf{c} &= \mathbf{C}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}) \\
&= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b})
\end{aligned} \tag{20}$$

Substituting back $f(\mathbf{x})$ into (18) we obtain

$$
\begin{aligned}
\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) &= \frac{1}{(2\pi)^D}\frac{1}{|\mathbf{AB}|^{\frac{1}{2}}}\exp\{f(\mathbf{x})\} \\
&= \frac{1}{(2\pi)^D}\frac{1}{|\mathbf{AB}|^{\frac{1}{2}}}\exp\left\{f(\mathbf{x}) - \frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c} + \frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c}\right\} \\
&= \frac{1}{(2\pi)^D}\frac{1}{|\mathbf{AB}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{x}^\top\mathbf{C}^{-1}\mathbf{x} + \mathbf{x}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c}\right. \\
&\quad \left. + \frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= \frac{1}{(2\pi)^D}\frac{1}{|\mathbf{AB}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{c})^\top\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c}) + \frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{|\mathbf{C}|^{\frac{1}{2}}}{|\mathbf{AB}|^{\frac{1}{2}}}\exp\left\{+\frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&\quad \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\mathbf{C}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{c})^\top\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c})\right\} \\
&= K^{-1}\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\mathbf{C}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{c})^\top\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c})\right\} \\
&= K^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c},\mathbf{C})
\end{aligned}
$$

2. Using the *matrix inversion lemma*, also known as the the Woodbury, Sherman & Morrison formula:

$$(\mathbf{Z} + \mathbf{UWV}^\top)^{-1} = \mathbf{Z}^{-1} - \mathbf{Z}^{-1}\mathbf{U}(\mathbf{W}^{-1} + \mathbf{V}^\top\mathbf{Z}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{Z}^{-1} \tag{21}$$

Proof that $\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A} - \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$

$$
\begin{aligned}
\mathbf{C} &= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = (\mathbf{Z} + \mathbf{UWV}^\top) = (\mathbf{A}^{-1} + \mathbf{IB}^{-1}\mathbf{I})^{-1} \\
&= (\mathbf{A}^{-1})^{-1} - (\mathbf{A}^{-1})^{-1}[(\mathbf{B}^{-1})^{-1} + \mathbf{I}(\mathbf{A}^{-1})^{-1}\mathbf{I}]^{-1}(\mathbf{A}^{-1})^{-1} \\
&= \mathbf{A} - \mathbf{A}(\mathbf{B} + \mathbf{IAI})^{-1}\mathbf{IA} \\
&= \mathbf{A} - \mathbf{A}(\mathbf{B} + \mathbf{A})^{-1}\mathbf{A} \\
&= \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}
\end{aligned} \tag{22}
$$

and by applying the same process

$$
\begin{aligned}
\mathbf{C} &= (\mathbf{B}^{-1} + \mathbf{A}^{-1})^{-1} = (\mathbf{Z} + \mathbf{UWV}^\top)^{-1} = (\mathbf{B}^{-1} + \mathbf{IA}^{-1}\mathbf{I})^{-1} \\
&= \mathbf{B} - \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} \\
&= \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} \\
\mathbf{C} &= (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}
\end{aligned} \tag{23}
$$

3. Show that

$$K^{-1} = (2\pi)^{-D/2}|\mathbf{A} + \mathbf{B}|^{-1/2}\exp\big(-\frac{1}{2}(\mathbf{a} - \mathbf{b})^\top(\mathbf{A} + \mathbf{B})^{-1}(\mathbf{a} - \mathbf{b})\big) \tag{24}$$

$$
\begin{aligned}
K^{-1} &= (2\pi)^{-\frac{D}{2}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{|\mathbf{AB}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}\mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= (2\pi)^{-\frac{D}{2}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{|\mathbf{AB}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}\left[(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}-\mathbf{B}^{-1}\mathbf{b})\right]^\top(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= (2\pi)^{-\frac{D}{2}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{|\mathbf{AB}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}(\mathbf{A}^{-1}\mathbf{a}-\mathbf{B}^{-1}\mathbf{b})^\top(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= (2\pi)^{-\frac{D}{2}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{|\mathbf{AB}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}-\mathbf{b}^\top\mathbf{B}^{-1})(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\} \\
&= (2\pi)^{-\frac{D}{2}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{|\mathbf{AB}|^{-\frac{1}{2}}} \exp\left\{\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}-\mathbf{b}^\top\mathbf{B}^{-1})[\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}](\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right. \\
&\qquad \left.-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right\}
\end{aligned}
$$

then developing the exponential term expanding the product

$$
\begin{aligned}
&= \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} + \frac{1}{2}\mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} - \frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{AB}^{-1}\mathbf{b} + \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{a} + \frac{1}{2}\mathbf{B}^{-1}\mathbf{AB}^{-1}\mathbf{b} \\
&\qquad - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{AB}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b} \\
&= -\frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} + \frac{1}{2}\mathbf{b}^\top(\mathbf{B}^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1})\mathbf{b} \\
&\qquad + \mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} - \mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{AB}^{-1}\mathbf{b}
\end{aligned}
\tag{25}
$$

applying the results from (23)

$$
\begin{aligned}
&\mathbf{B}^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1} \\
&\quad = \mathbf{B}^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1}\mathbf{A}(\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1})\mathbf{AB}^{-1} - \mathbf{B}^{-1} \\
&\quad = \mathbf{B}^{-1}\mathbf{AB}^{-1} - \mathbf{B}^{-1}\mathbf{AB}^{-1} + \mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} - \mathbf{B}^{-1} \\
&\quad = \mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} - \mathbf{B}^{-1}
\end{aligned}
\tag{26}
$$

and a second application gives

$$
\begin{aligned}
&\mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} - \mathbf{B}^{-1} \\
&\quad = (\mathbf{A}+\mathbf{B})^{-1}
\end{aligned}
\tag{27}
$$

and substituting back

$$
\begin{aligned}
-\frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} + \frac{1}{2}\mathbf{b}^\top(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{B}^{-1} - \mathbf{B}^{-1})\mathbf{b} \\
+ \mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} - \mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{b} \\
= -\frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} + \frac{1}{2}\mathbf{b}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} - \mathbf{a}^\top[(\mathbf{A}+\mathbf{B})^{-1} - \mathbf{B}^{-1}]\mathbf{b} \\
= -\frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} + \frac{1}{2}\mathbf{b}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} - \mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} + \mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} \\
= -\frac{1}{2}\mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{a} - \frac{1}{2}\mathbf{b}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} + \mathbf{b}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} - \mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} + \mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} \\
= (\mathbf{a}-\mathbf{b})^\top(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{a}-\mathbf{b}) - \mathbf{a}^\top(\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} + \mathbf{a}^\top\mathbf{B}^{-1}\mathbf{b} \quad (28)
\end{aligned}
$$

Problem 5. Tossing a biased coin with probability that it comes up heads is μ.

1. We toss the coin 3 times and it all comes up with heads. How likely is that in the next toss, the coin comes up with head according to MLE?

   We can model a single coin flip with a Bernoulli distribution where 1 means heads and 0 means tails

   $$X_i \sim \mathrm{Ber}(x|\mu) \begin{cases} \mu & x = 1 \\ 1-\mu & x = 0 \end{cases}$$

   so that the mean according to the MLE for $\mathcal{X} = \{X_1 = 1, X_2 = 1, X_3 = 1\}$ is given by

   $$\mu_{\mathrm{ML}} = \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i = 1$$

   We can then use μ_{ML} to predict the probability that the 4th coin toss will be head as follows

   $$p(X_4 = 1|\mu_{\mathrm{ML}}) = \mu_{\mathrm{ML}} = 1$$

2. Suppose that the prior $\mu \sim \mathrm{Beta}(\mu|a,b)$. What is the probability that the coin comes up with head in the 4th toss?

   First we need to compute the posterior mean

   $$
   \begin{aligned}
   p(\mu|\mathcal{X}) &= p(\mathcal{X}|\mu)p(\mu) \\
   &= p(X_1 = 1, X_2 = 1, X_3 = 1|\mu)p(\mu|a,b) \\
   &= \prod_{i=1}^{3} p(X_i = 1|\mu)p(\mu|a,b) \\
   &= \mu^3 \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
   &= \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a+2}(1-\mu)^{b-1}
   \end{aligned} \quad (29)
   $$

   concluding that it is also the probability of the 4th coin flip coming up heads.

3. Suppose that we observe m times that the coin lands heads and l times that it lands tails. Show that the posterior mean lies between the prior mean and μ_{MLE}.

   We can model the entire experiment as a Binomial distribution $X \sim \text{Bin}(x|m + l, \mu)$ so we have that the posterior mean is

$$\begin{aligned}
p(\mu|X) &= p(x|\mu)p(\mu) \\
&= \frac{(l+m)!}{l!m!}\mu^m(1-\mu)^l\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
&= \frac{(l+m)!}{l!m!}\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \\
&\propto \frac{\Gamma(a+b+m+l)}{\Gamma(a+m)+\Gamma(b+l)}\mu^{m+a-1}(1-\mu)^{l+b-1} \\
&= \text{Beta}(\mu|a+m, b+l)
\end{aligned} \tag{30}$$

   Because the posterior mean is a probability distribution, its value lies between 0 and 1, so it's less or equal than μ_{MLE}. We can then also note that the terms with μ of the prior are proportional to $a - 1$ and $b - 1$, while for the posterior they are $m + a - 1$ and $l + b - 1$, so the μ terms of the posterior grow faster than the prior. Moreover we know that the Gamma function is monotonically increasing for $a, b > 0$, so we can conclude that the posterior over the mean is greater or equal to the prior.

Problem 6$\star$. Derive mean, covariance, and mode of multivariate Student's t-distribution.