Assignment 1
Machine Learning 2, Spring 2017

Dana Kianfar
University of Amsterdam

January 6, 2018

Collaborators *Jose Gallego Posada, Mircea Mironenco, Jonas Köhler*

Resources See bibliography [1] [2] [3] [4].

Problem 1

We define our random variables as follows.

$$x \in \mathbb{R}^n, \quad p(x) = \mathcal{N}(x|\mu_x, \Sigma_x) = (2\pi)^{\frac{-n}{2}} \cdot |\Sigma_x|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(x-\mu_x)^T \Sigma_x^{-1}(x-\mu_x)\}$$

$$z \in \mathbb{R}^n, \quad p(z) = \mathcal{N}(z|\mu_z, \Sigma_z) = (2\pi)^{\frac{-n}{2}} \cdot |\Sigma_z|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(z-\mu_z)^T \Sigma_z^{-1}(z-\mu_z)\}$$

$$y = x + z$$

Mean As the expectation operation is linear, we can decompose it as follows.

$$\mathbb{E}[y] = \mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] = \mu_x + \mu_z$$

Covariance

$$\begin{aligned}
\Sigma_y &= \mathbb{E}[(y - \mathbb{E}[y])^2] = \mathbb{E}[(x+z-\mathbb{E}[x+z])^2] = \mathbb{E}[(x+z-\mu_x-\mu_z)^2] \\
&= \mathbb{E}[\underbrace{x^2}_{A} + \underbrace{z^2}_{B} + 2xz \underbrace{-2x\mu_x}_{A} - 2x\mu_z - 2z\mu_x \underbrace{-2z\mu_z}_{B} + \underbrace{\mu_x^2}_{A} + \underbrace{\mu_z^2}_{B} + 2\mu_x\mu_z] \\
&= \mathbb{E}[\underbrace{(x-\mu_x)^2}_{A} + \underbrace{(z-\mu_z)^2}_{B} + 2xz - 2x\mu_z - 2z\mu_x + 2\mu_x\mu_z] \\
&= \mathbb{E}[(x-\mu_x)^2 + (z-\mu_z)^2 + 2\mu_x(\mu_z - z) - 2\mu_x(\mu_z - z)] \\
&= \mathbb{E}[(x-\mu_x)^2 + (z-\mu_z)^2 + 2(\mu_z - z)(\mu_x - x)] = \Sigma_x + \Sigma_z + 2\mathbb{E}[(\mu_z - z)(\mu_x - x)] \\
\Sigma_y &= \Sigma_x + \Sigma_z + 2\text{Cov}[x, z]
\end{aligned}$$

We know that if $x \perp\!\!\!\perp z \rightarrow \text{Cov}[x, z] = 0$. Therefore, if x and z are independent, we have $\Sigma_y = \Sigma_x + \Sigma_z$.

Problem 2

$$\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T \in \mathbb{R}^{N \times D}, \quad \mathbf{x}_i \in \mathbb{R}^D, \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma), \quad \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \Sigma_0)$$

1.

$$p(\mathbf{x}|\boldsymbol{\mu},\Sigma) \overset{iid}{=} \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N} (2\pi)^{\frac{-D}{2}} \cdot |\Sigma|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\}$$

2. We can derive the posterior as follows. Note that the denominator is a constant term with respect to $\boldsymbol{\mu}$, and can be approximated up to a constant as follows.

$$p(\boldsymbol{\mu}|\mathcal{X},\Sigma,\boldsymbol{\mu}_0,\Sigma_0) = \frac{p(\mathcal{X}|\boldsymbol{\mu},\Sigma)\cdot p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\Sigma_0)}{p(\mathcal{X}|\Sigma,\boldsymbol{\mu}_0,\Sigma_0)} = \frac{p(\mathcal{X}|\boldsymbol{\mu},\Sigma)\cdot p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\Sigma_0)}{\int_{\boldsymbol{\mu}'} p(\mathcal{X}|\boldsymbol{\mu}',\Sigma)\cdot p(\boldsymbol{\mu}'|\boldsymbol{\mu}_0,\Sigma_0)\,d\boldsymbol{\mu}'}$$

$$\propto p(\mathcal{X}|\boldsymbol{\mu},\Sigma)\cdot p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\Sigma_0)$$

$$= \prod_{n=1}^{N} (2\pi)^{\frac{-D}{2}} \cdot |\Sigma|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\}$$

$$\times (2\pi)^{\frac{-D}{2}} \cdot |\Sigma_0|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\}$$

$$= (2\pi)^{\cancel{\frac{-ND-D}{2}}^{const}} \cdot |\Sigma^N\Sigma_0|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0))\}$$

3. Following our results from the previous question, we can expand the exponential term of the posterior (denoted here as A) as follows.

$$A = \exp\{\frac{-1}{2}(\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0))\}$$

$$= \exp\{\frac{-1}{2}(\sum_{n=1}^{N}(\mathbf{x}_n^T\Sigma^{-1}\mathbf{x}_n - \mathbf{x}_n^T\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma^{-1}\mathbf{x}_n + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0)\}$$

For any symmetric matrix $\mathbf{B}$, these two scalar results are equal $\mathbf{a_1^T B a_2} = \mathbf{a_2^T B a_1}$. Throughout this report, we use the symmetry of the covariance matrix. Therefore we can simplify the exponential term further.

$$A = \exp\{\frac{-1}{2}(\sum_{n=1}^{N}(\mathbf{x}_n^T\Sigma^{-1}\mathbf{x}_n - 2\boldsymbol{\mu}^T\Sigma^{-1}\mathbf{x}_n + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0)\}$$

$$= \exp\{\frac{-1}{2}(\cancel{\sum_{n=1}^{N}(\mathbf{x}_n^T\Sigma^{-1}\mathbf{x}_n)}^{const} \underbrace{-2\boldsymbol{\mu}^T\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n}_{1} \underbrace{-2\boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu}_0}_{1} + \underbrace{N\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}}_{2} + \underbrace{\boldsymbol{\mu}^T\Sigma_0^{-1}\boldsymbol{\mu}}_{2} + \cancel{(\boldsymbol{\mu}_0^T\Sigma_0^{-1}\boldsymbol{\mu}_0)}^{const}\}$$

$$= \exp\{\frac{-1}{2}(const \underbrace{-2\boldsymbol{\mu}^T(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0)}_{1} + \underbrace{\boldsymbol{\mu}^T(N\Sigma^{-1}+\Sigma_0^{-1})\boldsymbol{\mu}}_{2}\}$$

By substituting A into the posterior, we can see that it is a Gaussian function as the exponential term has quadratic, linear and constant terms (with respect to $\boldsymbol{\mu}$).

$$p(\boldsymbol{\mu}|\mathcal{X},\Sigma,\boldsymbol{\mu}_0,\Sigma_0) \propto |\Sigma^N\Sigma_0|^{\frac{-1}{2}} \cdot \exp\{\frac{-1}{2}(const + -2\boldsymbol{\mu}^T(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0) + \boldsymbol{\mu}^T(N\Sigma^{-1}+\Sigma_0^{-1})\boldsymbol{\mu})\}$$

To derive the parameters of our Gaussian posterior, we will further expand the exponential term (which we again denote as A). For simplicity, we ignore the coefficient $\frac{-1}{2}$ and the constant term *const* (and all resulting future constants) as they do not affect our results. Furthermore, we denote $\Lambda = N\Sigma^{-1} + \Sigma_0^{-1}$ and $\mathbf{B} = \Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0$.

$$A = \exp\{\boldsymbol{\mu}^T\Lambda\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\mathbf{B}\} = \exp\{\boldsymbol{\mu}^T\Lambda\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{B} - \boldsymbol{\mu}^T\mathbf{B}\} = \exp\{\boldsymbol{\mu}^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}) - \boldsymbol{\mu}^T\mathbf{B}\}$$

We add two redundant terms denoted by 1

$$= \exp\{\boldsymbol{\mu}^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}) - \underbrace{\mathbf{B}^T\Lambda^{-1}\mathbf{B} + \mathbf{B}^T\Lambda^{-1}\mathbf{B}}_{1} - \boldsymbol{\mu}^T\mathbf{B}\} = \exp\{\boldsymbol{\mu}^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}) - \cancel{\mathbf{B}^T\Lambda^{-1}\mathbf{B}}^{const} + \mathbf{B}^T\Lambda^{-1}\mathbf{B} - \mathbf{B}^T\boldsymbol{\mu}\}$$

$$= \exp\{\boldsymbol{\mu}^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}) - \mathbf{B}^T(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B})\}$$

We introduce a redundant matrix multiplication $\Lambda^{-1}\Lambda$

$$= \exp\{\boldsymbol{\mu}^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}) - \mathbf{B}^T\Lambda^{-1}\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B})\} = \exp\{(\boldsymbol{\mu}^T - \mathbf{B}^T\Lambda^{-1})\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}))\}$$

$$= \exp\{(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B})^T\Lambda(\boldsymbol{\mu} - \Lambda^{-1}\mathbf{B}))\}$$

We can therefore express μ_N and Σ_N as follows.

$$\Sigma_N = \Lambda^{-1} = (N\Sigma^{-1} + \Sigma_0^{-1})^{-1}, \quad \mu_N = \Lambda^{-1}\mathbf{B} = \Sigma_N(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x}_n + \Sigma_0^{-1}\boldsymbol{\mu}_0)$$

4. Given that our posterior is a Gaussian, it is symmetric and unimodal, the mode coincides with the mean. Our maximum a posteriori (MAP) estimator is expressed as follows.

$$\boldsymbol{\mu}_{MAP} = \underset{\boldsymbol{\mu}}{argmax} \quad p(\boldsymbol{\mu}|\mathcal{X}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0) = \boldsymbol{\mu}_N$$

We can derive this result by taking the derivative of the log-posterior and solving for $\boldsymbol{\mu}$ as follows.

$$\log p(\boldsymbol{\mu}|\mathcal{X}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0) \propto \log\exp\{\frac{-1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^T\Sigma_N^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)\}$$

$$\propto \boldsymbol{\mu}^T\Sigma_N^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\Sigma_N^{-1}\boldsymbol{\mu}_N + \boldsymbol{\mu}_N^T\Sigma_N^{-1}\boldsymbol{\mu}_N$$

$$\frac{\partial \log p(\boldsymbol{\mu}|\mathcal{X}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0)}{\partial \boldsymbol{\mu}} = 2\Sigma_N^{-1}\boldsymbol{\mu} - 2\Sigma_N^{-1}\boldsymbol{\mu}_N$$

$$\frac{\partial \log p(\boldsymbol{\mu}|\mathcal{X}, \Sigma, \boldsymbol{\mu}_0, \Sigma_0)}{\partial \boldsymbol{\mu}} = 0 \rightarrow \Sigma_N^{-1}\boldsymbol{\mu} = \Sigma_N^{-1}\boldsymbol{\mu}_N \rightarrow \boldsymbol{\mu} = \mu_N$$

5. We will denote $\sum_{n=1}^{N}\mathbf{x}_n$ as $N \cdot \mu_{MLE}$, where $\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N}x_n$. Before updating our posterior with new datapoints, our parameters are as follows.

$$\sigma_N^2 = (\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2})^{-1} = \frac{\sigma^2\sigma_0^2}{\sigma^2 + N\sigma_0^2}, \quad \mu_N = \sigma_N^2(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{n=1}^{N}x_n}{\sigma^2}) = \sigma_N^2(\frac{\mu_0}{\sigma_0^2} + \frac{(N)\mu_{MLE}}{\sigma^2})$$

While updating, the current posterior becomes the new prior. The update rule for the variance is defined as follows.

$$\frac{1}{\sigma_{N+1}^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_N^2} + \frac{1}{\sigma^2} \rightarrow \sigma_{N+1}^2 = (\frac{1}{\sigma_N^2} + \frac{1}{\sigma^2})^{-1}$$

We can decompose μ_N to a convex combination between μ_0 and μ_{MLE}.

$$\mu_N = \sigma_N^2(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{n=1}^{N}x_n}{\sigma^2}) = \sigma_N^2(\frac{\sigma^2\mu_0 + \sigma_0^2\sum_{n=1}^{N}x_n}{\sigma_0^2\sigma^2}) = (\frac{\cancel{\sigma^2\sigma_0^2}}{\sigma^2 + N\sigma_0^2})(\frac{\sigma^2\mu_0 + \sigma_0^2\sum_{n=1}^{N}x_n}{\cancel{\sigma_0^2\sigma^2}})$$

$$= \frac{\sigma_0^2\sum_{n=1}^{N}x_n}{\sigma^2 + N\sigma_0^2} + \frac{\sigma^2\mu_0}{\sigma^2 + N\sigma_0^2} = \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}(\mu_{MLE}) + \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}(\mu_0)$$

After observing the $N + 1^{th}$ datapoint, we can write the update rule for the mean in a similar manner.

$$\mu_{N+1} = \frac{\sigma_N^2}{\sigma^2 + \sigma_N^2}(x_{N+1}) + \frac{\sigma^2}{\sigma^2 + \sigma_N^2}(\mu_N)$$

6. Following the methodology from Q2.5, we can write the update rules directly from the posterior distribution.

$$p(\mu|x_1, x_2, \cdots, x_N) \propto \exp\{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 + \frac{-1}{2\sigma_0^2}(\mu - \mu_0)^2\}$$

$$= \exp\{\frac{-1}{2\sigma^2}\cancel{\sum_{n=1}^{N}x_n^2}^{const} - \frac{N}{2\sigma^2}\mu^2 + \frac{1}{\sigma^2}\mu\sum_{n=1}^{N}x_n - \frac{\mu^2}{2\sigma_0^2} - \cancel{\frac{\mu_0^2}{2\sigma_0^2}}^{const} + \frac{\mu\mu_0}{\sigma_0^2}\}$$

$$= \exp\{const - \underbrace{\mu^2(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2})}_{A_1} + \underbrace{\mu(\frac{\sum_{n=1}^N x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})}_{A_2}\}$$

Note that the posterior may also be expressed as $\mathcal{N}(\mu|\mu_N, \sigma_N)$. The exponential term of posterior (denoted here as B) can also be expressed as follows.

$$B := \exp\{\frac{-1}{2\sigma_N^2}(\mu - \mu_N)^2\} = \exp\{\underbrace{\frac{-1}{2\sigma_N^2}\mu^2}_{B_1} + \underbrace{\frac{1}{\sigma_N^2}\mu\mu_N)}_{B_2} - \frac{1}{2\sigma_N^2}\mu_N^2\}$$

By matching the quadratic terms A_1 and B_1, and the linear terms A_2 and B_2 with respect to μ, we can obtain the update rules. For brevity, we re-use our derivations from previous questions.

$$A_1 \& B_1 : \frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2} = \frac{1}{2\sigma_N^2} \rightarrow \frac{N\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2} = \frac{1}{\sigma_N^2} \rightarrow \sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

$$\sigma_{N+1}^2 = (\frac{1}{\sigma_N^2} + \frac{1}{\sigma^2})^{-1}$$

$$A_2 \& B_2 : \frac{\sum_{n=1}^N x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} = \frac{\mu_N}{\sigma_N^2} \rightarrow \mu_N = (\sigma_N^2)^{-1}\frac{\sum_{n=1}^N x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$$

$$\mu_{N+1} = \frac{\sigma_N^2}{\sigma^2 + \sigma_N^2}(x_{N+1}) + \frac{\sigma^2}{\sigma^2 + \sigma_N^2}(\mu_N)$$

Problem 3

1. We simplify the demonstration, as the procedure is identical to completing the square shown in q2.3. For this reason, we omit constant terms (including the $\frac{-1}{2}$ coefficient in the exponential term).

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) \propto \exp\{(x-a)^T A^{-1}(x-a) + (x-b)^T B^{-1}(x-b)\}$$

$$= \exp\{x^T A^{-1}x - 2x^T A^{-1}a + a^T A^{-1}a + x^T B^{-1}x - 2x^T B^{-1}b + b^T B^{-1}b\}$$

$$= \exp\{x^T(A^{-1} + B^{-1})x - 2x^T(A^{-1}a + B^{-1}b) + \cancel{a^T A^{-1}a + b^T B^{-1}b}\}$$ *const*

We can clearly see quadratic, linear and constant terms with respect to x. As this procedure is identical to Question 2.3, we proceed to present the final derivations below. The constant term is denoted by c.

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) \propto \exp\{(x-(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b))^T(A^{-1}+B^{-1})(x-(A^{-1}+B^{-1})^{-1}(A^{-1}a+B^{-1}b))+c\}$$

Thus we have $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$.

2. We take $\mathbf{Z} = \mathbf{A^{-1}}, \mathbf{U} = \mathbf{V} = \mathbb{I}, \mathbf{W} = \mathbf{B^{-1}}$.

$$\mathbf{C} = (\mathbf{A^{-1}} + \mathbb{I}\mathbf{B^{-1}}\mathbb{I})^{\mathbf{-1}} = \mathbf{A} - \mathbf{A}\mathbb{I}(\mathbf{B} + \mathbb{I}\mathbf{A}\mathbb{I})^{\mathbf{-1}}\mathbb{I}\mathbf{A} = \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B})^{\mathbf{-1}}\mathbf{A}$$

By reassigning our terms such that $\mathbf{Z} = \mathbf{B^{-1}}, \mathbf{W} = \mathbf{A^{-1}}$, we can obtain the second result.

$$\mathbf{C} = (\mathbf{A^{-1}} + \mathbb{I}\mathbf{B^{-1}}\mathbb{I})^{\mathbf{-1}} = \mathbf{B} - \mathbf{B}\mathbb{I}(\mathbf{B} + \mathbb{I}\mathbf{A}\mathbb{I})^{\mathbf{-1}}\mathbb{I}\mathbf{B} = \mathbf{B} - \mathbf{B}(\mathbf{A} + \mathbf{B})^{\mathbf{-1}}\mathbf{B}$$

3. We use Eqns (21, 163, 165) from the Matrix Cookbook in this section. On each side of the given formula we have the following.

$$LHS : \mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = \cancel{(2\pi)^{\frac{-D}{2}}}|AB|^{\frac{-1}{2}}\exp\{\frac{-1}{2}(x-a)^T A^{-1}(x-a) - \frac{1}{2}(x-b)^T B^{-1}(x-b)\}$$

$$RHS : \mathcal{K}^{-1}\mathcal{N}(x|c, C) = \cancel{(2\pi)^{\frac{-D}{2}}}|C|^{\frac{-1}{2}}|A+B|^{\frac{-1}{2}}\exp\{\frac{-1}{2}(a-b)^T(A+B)^{-1}(a-b) - \frac{1}{2}(x-c)^T C^{-1}(x-c)\}$$

We can prove that the determinant terms are equal using Equation (165) from the Matrix Cookbook.

$$|C|\cdot|A+B| = \frac{|A+B|}{|A^{-1}+B^{-1}|} = \frac{|A+B|}{|A^{-1}(A+B)B^{-1}|} = \frac{1}{|A^{-1}B^{-1}|} = |AB|$$

To prove the equality of the exponential term, we will attempt to equate every term on the left hand size (LHS) and the right hand side (RHS). Once we have removed all equal corresponding terms from each side, we will have proven equality if both sides are equal to zero.

Recall that we have previously shown that all terms dependent on x cancel eachother out. Therefore, we are left with the following terms on each side. For simplicity we omit the coefficient $-\frac{1}{2}$. We use Eqn (163) from the Matrix Cookbook to simplify C.

$$LHS : \exp\{a^T A^{-1} a + b^T B^{-1} b\}$$

$$RHS : \exp\{(a-b)^T (A+B)^{-1}(a-b) + c^T C^{-1} c\}$$

$$c^T C^{-1} c = (A^{-1}a + B^{-1}b)^T C \cancel{C^{-1} C} (A^{-1}a + B^{-1}b) = a^T A^{-1} C A^{-1} a + b^T B^{-1} C B^{-1} b + a^T A^{-1} C B^{-1} b + b^T B^{-1} C A^{-1} a$$

$$= a^T A^{-1}(A - A(A+B)^{-1}A)A^{-1}a + B^T B^{-1}(B - B(A+B)^{-1}B)B^{-1}b$$

$$+ a^T A^{-1} A (A+B)^{-1} B B^{-1} b + b^T B^{-1} B (A+B)^{-1} A A^{-1} a$$

$$= a^T A^{-1} a + b^T B^{-1} b - a^T (A+B)^{-1} a - b^T (A+B)^{-1} b + b^T (A+B)^{-1} a + b^T (A+B)^{-1} b$$

$$= a^T A^{-1} a + b^T B^{-1} b - (a-b)^T (A+B)^{-1}(a+b)$$

$$LHS : \exp\{\cancel{a^T A^{-1} a + b^T B^{-1} b}\} = 0$$

$$RHS : \exp\{(a-b)^T (A+B)^{-1}(a-b) + (a-b)^T (A+B)^{-1}(a+b) + \cancel{a^T A^{-1} a + b^T B^{-1} b}\} = 0$$

The two sides are empty, therefore the equality has been proved, and we have $\mathcal{K}^{-1} = (2\pi)^{\frac{-D}{2}} |A+B|^{\frac{-1}{2}} \exp\{\frac{-1}{2}(a-b)^T (A+B)^{-1}(a-b)\}$.

Problem 4

1. First we derive MLE estimator for μ. We denote $m = \sum_{n=1}^{N} x_n$, and $l = \sum_{n=1}^{N} 1 - x_n$ such that $m + l = N$. $\mathcal{L}$ is the likelihood function of $\mathcal{X}$.

$$\mathcal{X} = (x_1, x_2, x_3)^T, \quad \mathcal{L}(\mathcal{X}, \mu, m, l) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{(1-x_n)}, \quad \log \mathcal{L}(\mathcal{X}, \mu, m, l) = \sum_{n=1}^{N} x_n \log \mu - (1-x_n)\log(1-\mu)$$

$$\frac{\partial \log \mathcal{L}}{\partial \mu} = \frac{1}{\mu}\sum_{n=1}^{N} x_n - \frac{1}{1-\mu}\sum_{n=1}^{N} 1 - x_n = 0 \rightarrow \frac{m}{\mu} = \frac{N-m}{1-\mu} \rightarrow \mu = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{m+l}$$

   The maximum likelihood estimator for the probability of getting heads is $\mu_{MLE} = \frac{3}{3+0} = 1$, given $m = 3$, $l = 0$. Therefore, by relying solely on our dataset, an MLE model expects the next coin flip to also be heads with probability 1.

2. In a Bayesian treatment, we have our posterior defined as follows. For simplicity, we omit the normalization constants (the full posterior is a beta distribution given in Bishop 2.18).

$$p(\mu|\mathcal{X}, a, b) = \frac{p(\mathcal{X}|\mu)p(\mu|a,b)}{\int_{\mu'} p(\mathcal{X}|\mu')p(\mu'|a,b)d\mu'} \propto p(\mathcal{X}|\mu)p(\mu|a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

$$\log p(\mu|\mathcal{X}, a, b) = (m+a-1)\log\mu - (l+b-1)\log(1-\mu)$$

$$\frac{\partial \log p(\mu|\mathcal{X}, a, b)}{\partial \mu} = \frac{m+a-1}{\mu} - \frac{l+b-1}{1-\mu} = 0 \rightarrow \mu_{MAP} = \frac{m+a-1}{m+l+a+b-2}$$

   In our setting, the MAP estimator is $\mu_{MAP} = \frac{a+2}{a+b+1}$. To predict whether the 4th flip will be heads, we can use the predictive distribution given by Bishop (2.20), where we obtain $p(x=1|\mathcal{X}) = \frac{m+a}{m+a+l+b} = \frac{a+3}{a+b+3}$.

3. The posterior mean is $\frac{m+l}{m+l+a+b}$, the MLE mean is $\frac{m}{m+l}$ and the prior mean is $\frac{a}{a+b}$. We can demonstrate that the posterior mean is a convex combination of the prior mean and the MLE mean.

$$\frac{m+a}{m+a+l+b} = \frac{m}{m+a+l+b} + \frac{a}{m+a+l+b} = \frac{m+l}{m+a+l+b}\left(\frac{m}{m+l}\right) + \frac{a+b}{m+a+l+b}\left(\frac{a}{a+b}\right)$$

Problem 5

To calculate the expectation, we use Bishop (2.226).

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

Poisson

$$p(x|\lambda) \sim Poisson(x|\lambda) = \frac{\lambda^x e^\lambda}{x!} = \frac{1}{x!}\exp\{x\log\lambda - \lambda\}$$

$$h(x) = \frac{1}{x!}, \quad \eta = log\lambda, \quad g(\eta) = \exp\{-e^\eta\}, \quad u(x) = x$$

$$\mathbb{E}[x] = -\nabla_\eta \log g(\eta) = -\frac{\partial(-e^\eta)}{\partial\eta} = e^\eta = e^{\log\lambda} = \lambda$$

Gamma

$$p(x|a,b) \sim Gam(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx} = \frac{b^a}{\Gamma(a)}x^{-1}\exp\{a\log x - bx\}$$

$$h(x) = \frac{1}{x}, \quad \eta = [a,b]^T, \quad g(\eta) = \frac{b^a}{\Gamma(a)}, \quad u(x) = [\log x, -x]^T$$

To derive the expectation, we must take the partial derivative with respect to η_2, corresponding to $-u_2(x)$, which represents x.

$$\mathbb{E}[x] = -\nabla_{\eta_2}\log g(\eta_2) = \frac{\partial(a\log b - \log(\Gamma(a))}{\partial b} = \frac{a}{b}$$

Problem 6*

Mean We use Bishop's formulation of the multivariate t-distribution in eqn (2.161), expressed as follows.

$$st(\mathbf{x}|\boldsymbol{\mu},\Lambda,\nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1})Gam(\eta|\nu/2,\nu/2)d\eta$$

$$\mathbb{E}[x] = \int xp(x)\,dx = \int x\int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1})Gam(\eta|\nu/2,\nu/2)d\eta dx = \int\int_0^\infty x\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1})Gam(\eta|\nu/2,\nu/2)dxd\eta$$

$$= \int_0^\infty \boldsymbol{\mu}Gam(\eta|\nu/2,\nu/2)d\eta = \boldsymbol{\mu}\int_0^\infty Gam(\eta|\nu/2,\nu/2)d\eta = \boldsymbol{\mu}$$

Covariance The following result holds if $\nu > 2$.

$$\mathrm{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1})Gam(\eta|\nu/2,\nu/2)d\eta dx$$

$$= \int\int_0^\infty (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\Lambda)^{-1})Gam(\eta|\nu/2,\nu/2)dxd\eta$$

$$= \int_0^\infty (\eta\Lambda)^{-1}Gam(\eta|\nu/2,\nu/2)d\eta$$

$$= \Lambda^{-1}\int_0^\infty \frac{1}{\eta}Gam(\eta|\nu/2,\nu/2)d\eta$$

$$= \Lambda^{-1}\int_0^\infty \frac{1}{\eta}\cdot\frac{(\nu/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)}\eta^{\frac{\nu}{2}-1}e^{-\frac{\nu}{2}\eta}d\eta$$

$$= \Lambda^{-1}\int_0^\infty \frac{(\nu/2)^{\frac{\nu}{2}-1}(\nu/2)}{(\nu/2-1)\Gamma(\nu/2-1)}\eta^{\frac{\nu}{2}-2}e^{-\frac{\nu}{2}\eta}d\eta$$

$$= \Lambda^{-1} \frac{(\nu/2)}{(\nu/2-1)} \int_0^\infty \frac{(\nu/2)^{\frac{\nu}{2}-1}}{\Gamma(\nu/2-1)} \eta^{\frac{\nu}{2}-2} e^{-\frac{\nu}{2}\eta} d\eta = \Lambda^{-1} \frac{(\nu/2)}{(\nu/2-1)} = \Lambda^{-1} \frac{\nu}{\nu-2}$$

Mode The integral formulation can be evaluated as follows (given by Bishop (2.162)).

$$st(\mathbf{x}|\boldsymbol{\mu}, \Lambda, \nu) = \frac{\Gamma(D/2+\nu/2)}{\Gamma(\mu/2)} \frac{|\Lambda|^{1/2}}{(\pi\nu)^{D/2}} [1 + \frac{(\mathbf{x}-\boldsymbol{\mu})^T \Lambda (\mathbf{x}-\boldsymbol{\mu})}{\nu}]^{-D/2-\nu/2}$$

The pmf of $st(\mathbf{x}|\boldsymbol{\mu}, \Lambda, \nu)$ is monotonously decreasing in the mahalanobis distance term $(\mathbf{x}-\boldsymbol{\mu})^T \Lambda (\mathbf{x}-\boldsymbol{\mu})$. Therefore the mode is where this distance is zero, which is when $\mathbf{x} = \boldsymbol{\mu}$.

References

[1] Kimiyoshi Oshikoji. Sequential estimation/bayesian inference for gaussian distribution, 2010.

[2] Kevin Murphy. Parameter estimation for univariate gaussian distributions, 2006.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, nov 2012.