Chapter 6 Solutions

Exercise 6.1

a)

The marginal distributions are obtained by summing the probabilies over all the values of the variable being marginalized. Thus, to obtain $p(x)$ we sum over columns (i.e., over the values corresponding to different y):

$$p(x_1) = P(X = x_1) = P(X = x_1, Y = y_1) + P(X = x_1, Y = y_2) + P(X = x_1, Y = y_3) = 0.01 + 0.05 +$$
$$p(x_2) = P(X = x_2) = P(X = x_2, Y = y_1) + P(X = x_2, Y = y_2) + P(X = x_2, Y = y_3) = 0.02 + 0.1 +$$
$$p(x_3) = P(X = x_3) = P(X = x_3, Y = y_1) + P(X = x_3, Y = y_2) + P(X = x_3, Y = y_3) = 0.03 + 0.05 +$$
$$p(x_4) = P(X = x_4) = P(X = x_4, Y = y_1) + P(X = x_4, Y = y_2) + P(X = x_4, Y = y_3) = 0.1 + 0.07 +$$
$$p(x_5) = P(X = x_5) = P(X = x_5, Y = y_1) + P(X = x_5, Y = y_2) + P(X = x_5, Y = y_3) = 0.1 + 0.2 + 0.$$

As a correctness check, note that this distribution satisfies the normalization condition, i.e. that sum of the probabilities is 1:

$\sum_{i=1}^{5} p(x_i) = 1$

The marginal distribution $p(y)$ can be obtained in a similar way, by summing the matrix rows:

$$p(y_1) = P(Y = y_1) = \sum_{i=1}^{5} P(X = x_i, Y = y_1) = 0.01 + 0.02 + 0.03 + 0.1 + 0.1 = 0.26$$

$$p(y_2) = P(Y = y_2) = \sum_{i=1}^{5} P(X = x_i, Y = y_2) = 0.05 + 0.1 + 0.05 + 0.07 + 0.2 = 0.47$$

$$p(y_3) = P(Y = y_3) = \sum_{i=1}^{5} P(X = x_i, Y = y_3) = 0.1 + 0.05 + 0.03 + 0.05 + 0.04 = 0.27$$

We can again check that the normalization condition is satisfied:

$\sum_{i=1}^{3} p(y_i) = 1$

b)

To determine conditional distributions we use the definition of the conditional probability:

$P(X = x, Y = y_1) = P(X = x|Y = y_1)P(Y = y_1) = p(x|Y = y_1)p(y_1).$

Thus,

$p(x_1|Y=y_1) = \frac{P(X=x_1,Y=y_1)}{p(y_1)} = \frac{0.01}{0.26} \approx 0.038$
$p(x_2|Y=y_1) = \frac{P(X=x_2,Y=y_1)}{p(y_1)} = \frac{0.02}{0.26} \approx 0.077$
$p(x_3|Y=y_1) = \frac{P(X=x_3,Y=y_1)}{p(y_1)} = \frac{0.03}{0.26} \approx 0.115$
$p(x_4|Y=y_1) = \frac{P(X=x_4,Y=y_1)}{p(y_1)} = \frac{0.1}{0.26} \approx 0.385$
$p(x_5|Y=y_1) = \frac{P(X=x_5,Y=y_1)}{p(y_1)} = \frac{0.1}{0.26} \approx 0.385$

Likewise the conditional distribution $p(y|X=x_3)$ is given by

$p(y_1|X=y_3) = \frac{P(X=x_3,Y=y_1)}{p(x_3)} = \frac{0.03}{0.11} \approx 0.273$
$p(y_2|X=y_3) = \frac{P(X=x_3,Y=y_2)}{p(x_3)} = \frac{0.05}{0.11} \approx 0.454$
$p(y_3|X=y_3) = \frac{P(X=x_3,Y=y_3)}{p(x_3)} = \frac{0.03}{0.11} \approx 0.273$

Exercise 6.2

a)

We can write the probability density of the two-dimensional distribution as

$$p(x,y) = 0.4\mathcal{N}\left(x,y|\begin{bmatrix}10\\2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right) + 0.6\mathcal{N}\left(x,y|\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}\right)$$

The marginal distribution of a weighted sum of distributions is given by the weighted sum of marginals, whereas the marginals of a bivariate normal distribution $\mathcal{N}(x,y|\mu,\boldsymbol{\Sigma})$ are obtained according to the rule

$\int \mathcal{N}(x,y|\mu,\boldsymbol{\Sigma})dy = \mathcal{N}(x|\mu_x,\Sigma_{xx}),$
$\int \mathcal{N}(x,y|\mu,\boldsymbol{\Sigma})dx = \mathcal{N}(y|\mu_y,\Sigma_{yy})$

Thus, the marginals of the distribution of interest are

$p(x) = 0.4\mathcal{N}(x|10,1) + 0.6\mathcal{N}(x|0,8.4),$
$p(y) = 0.4\mathcal{N}(x|2,1) + 0.6\mathcal{N}(x|0,1.7)$

b)

The mean of a weighted sum of two distributions is the weighted sum of their averages

$\mathbb{E}_X[x] = 0.4*10 + 0.6*0 = 4,$
$\mathbb{E}_Y[y] = 0.4*2 + 0.6*0 = 0.8$

The mode of a continuous distribution is a point where this distribution has a peak. It can be determined by solving the extremum condition for each of the marginal distributions:

$\frac{dp(x)}{dx} = 0,$
$\frac{dp(y)}{dy} = 0$

In the case of a mixture of normal distributions these equations are non-linear and can be solved only numerically. After finding all the solutions of these equations one has to verify for every solution that it is a peak rather than an inflection point, i.e. that at this point

$$\frac{d^2p(x)}{dx^2} < 0 \text{ or } \frac{d^2p(y)}{dy^2} < 0$$

The medians m_x, m_y can be determined from the conditions

$$\int_{-\infty}^{m} p(x)dx = \int_{m}^{+\infty} p(x)dx,$$
$$\int_{-\infty}^{m} p(y)dy = \int_{m}^{+\infty} p(y)dy$$

Again, these equations can be solved here only numerically.

c)

The mean of a two-dimensional distribution is a vector of means of the marginal distributions

$$\mu = \begin{bmatrix} 4 \\ 0.8 \end{bmatrix}$$

The mode of two dimensional distribution is obtained first by solving the extremum conditions

$$\frac{\partial p(x,y)}{\partial x} = 0, \frac{\partial p(x,y)}{\partial y} = 0$$

and then verifying for every solution that it is indeed a peak, i.e.

$$\frac{\partial^2 p(x,y)}{\partial x^2} < 0, \frac{\partial^2 p(x,y)}{\partial y^2} < 0,$$
$$\det\left(\begin{bmatrix} \frac{\partial^2 p(x,y)}{\partial x^2} & \frac{\partial^2 p(x,y)}{\partial x \partial y} \\ \frac{\partial^2 p(x,y)}{\partial x \partial y} & \frac{\partial^2 p(x,y)}{\partial y^2} \end{bmatrix}\right) > 0$$

Again, these squations can be solved only numerically.

Exercise 6.3

The conjugate prior to the Bernoulli distribution is the Beta distribution

$$p(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1},$$

where α, β are not necessarily integers and the normalization coefficient si the Beta function defined as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

The likelihood of observing data $\{x_1, x_2, \ldots, x_N\}$ is

$$p(x_1,\ldots,x_N|\mu) = \prod_{i=1}^N p(x_i|\mu) = \prod_{i=1}^N \mu^{x_i}(1-\mu)^{1-x_i} = \mu^{\sum_{i=1}^N x_i}(1-\mu)^{N-\sum_{i=1}^N x_i}$$

The posterior distribution is proportional to teh rpoduct of this likelihood with teh prior distribution (Bayes theorem):

$$p(\mu|x_1,\ldots,x_N) \propto p(x_1,\ldots,x_N|\mu)p(\mu|\alpha,\beta) \propto \mu^{\sum_{i=1}^N x_i+\alpha-1}(1-\mu)^{N-\sum_{i=1}^N x_i+\beta-1}$$

This is also a Beta distribution, i.e. our choice of the gonjugate prior was correct. The normalization constant is readily determined:

$$p(\mu|x_1,\ldots,x_N) = \frac{1}{\mathcal{B}(\sum_{i=1}^N x_i+\alpha-1, N-\sum_{i=1}^N x_i+\beta-1)}\mu^{\sum_{i=1}^N x_i+\alpha-1}(1-\mu)^{N-\sum_{i=1}^N x_i+\beta-1}$$

Exercise 6.4

The probabilities of picking a mango or an apple from teh first bag are given by

$$p(mango|1) = \frac{4}{6} = \frac{2}{3}$$
$$p(apple|1) = \frac{2}{6} = \frac{1}{3}$$

The probabilities of picking a mango or an apple from teh second bag are
$$p(mango|2) = \frac{4}{8} = \frac{1}{2}$$
$$p(apple|2) = \frac{4}{8} = \frac{1}{2}$$

The probability of picking the first or the second bag are equal to teh probabilities of head and tail respectively:

$$p(1) = 0.6,$$
$$p(2) = 0.4$$

We now can obtain the probability that the mango was picked from the second bag using Bayes' theorem:

$$p(2|mango) = \frac{p(mango|2)p(2)}{p(mango)} = \frac{p(mango|2)p(2)}{p(mango|1)p(1)+p(mango|2)p(2)} = \frac{\frac{1}{2}0.4}{\frac{2}{3}0.6+\frac{1}{2}0.4} = \frac{1}{3}$$

Exercise 6.5

a)

$\mathbf{x}_{t+1}$ is obtained from $\mathbf{x}_t$ by a linear transformation, $\mathbf{A}\mathbf{x}_t$ and adding a Gaussian random variabme $\mathbf{w}$. Initial distribution for $\mathbf{x}_0$ is a Gaussian distribution, a linear transformation of a Gaussian random variable is also a Gaussian random variable, whareas a sum of Gaussian random variables is a Gaussian random variable. Thus, the joint distribution $p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T)$ is also a Gaussian distribution.

b)

1)

Let $\mathbf{z} = \mathbf{A}\mathbf{x}_{t+1}$. Since this is a linear transformation of a Gaussian random variable, $\mathbf{x}_t \sim \mathcal{N}(\mu_t, \mathbf{\Sigma})$, then $\mathbf{z}$ is distributed as (see Eq. (6.88))

$$\mathbf{z} \sim \mathcal{N}(\mathbf{A}\mu_t, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T),$$

whereas the mean and the covariance of a sum of two Gaussian random variables are given by the sum of the means and the covariances of these variables, i.e.,

$$\mathbf{x}_{t+1} = \mathbf{z} + \mathbf{w} \sim \mathcal{N}(\mathbf{A}\mu_t, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q}),$$

That is

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q}).$$

2)

If we assume that $\mathbf{x}_{t+1}$ is fixed, then $\mathbf{y}_{t+1} = \mathbf{C}\mathbf{x}_{t+1} + \mathbf{v}$ follows the same distribution as $\mathbf{v}$, but with the mean shifted by $\mathbf{C}\mathbf{x}_{t+1}$, i.e.

$$p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathcal{N}(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{x}_{t+1}, \mathbf{R}).$$

The the joint probability is obtained as

$$p(\mathbf{y}_{t+1}, \mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{y}_1, \ldots, \mathbf{y}_t)p(\mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathcal{N}(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{x}_{t+1}, \mathbf{R})\mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\mathbf{\Sigma}$$

3)

Let us introduce temporary notation

$$\mu_{t+1} = \mathbf{A}\mu_t,$$
$$\mathbf{\Sigma}_{t+1} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q},$$
$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathcal{N}(\mu_{t+1}, \mathbf{\Sigma}_{t+1})$$

Then $\mathbf{y}_{t+1}$ is obtained in terms of the parameters of distribution $p(\mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t)$ following the same steps as question 1), with the result

$$p(\mathbf{y}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t) = \mathcal{N}(\mathbf{y}_{t+1}|\mathbf{C}\mu_{t+1}, \mathbf{C}\mathbf{\Sigma}_{t+1}\mathbf{C}^T + \mathbf{R}) = \mathcal{N}\left(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{A}\mu_t, \mathbf{C}(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q})\mathbf{C}^T + \mathbf{R}\right).$$

The required conditional distribution is then obtained as

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t, \mathbf{y}_{t+1}) = \frac{p(\mathbf{y}_{t+1}, \mathbf{x}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t)}{p(\mathbf{y}_{t+1}|\mathbf{y}_1, \ldots, \mathbf{y}_t)} = \frac{\mathcal{N}(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{x}_{t+1}, \mathbf{R})\mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q})}{\mathcal{N}\left(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{A}\mu_t, \mathbf{C}(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{Q})\mathbf{C}^T + \mathbf{R}\right)}$$

Exercise 6.6

The standard definition of variance is

$$\mathbb{V}_X[x] = \mathbb{E}_X[(x - \mu)^2],$$

where

$$\mu = \mathbb{E}_X[x].$$

Using the properties of average we can write:

$$\mathbb{V}_X[x] = \mathbb{E}_X[(x - \mu)^2] = \mathbb{E}_X[x^2 - 2x\mu + \mu^2] = \mathbb{E}_X[x^2] - \mathbb{E}_X[2x\mu] + \mathbb{E}_X[\mu^2] =$$
$$\mathbb{E}_X[x^2] - 2\mu\mathbb{E}_X[x] + \mu^2 = \mathbb{E}_X[x^2] - 2\mu^2 + \mu^2 = \mathbb{E}_X[x^2] - \mu^2$$

By substituting to this equation the definition of μ, we obtain the desired equation

$\mathbb{V}_X[x] = \mathbb{E}_X[(x-\mu)^2] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$

Exercise 6.7

Let is expand the square in the left-hand side of (6.45)

$$\frac{1}{N^2}\sum_{i,j=1}^N (x_i - x_j)^2 = \frac{1}{N^2}\sum_{i,j=1}^N (x_i^2 - 2x_i x_j + x_j^2) = \frac{1}{N^2}\sum_{i,j=1}^N x_i^2 - 2\frac{1}{N^2}\sum_{i,j=1}^N x_i x_j + \frac{1}{N^2}\sum_{i,j=1}^N x_j^2$$

We see that the first and the last term differ only by the summation index, i.e. they are identical:
$\frac{1}{N^2}\sum_{i,j=1}^N x_i^2 + \frac{1}{N^2}\sum_{i,j=1}^N x_j^2 = 2\frac{1}{N^2}\sum_{i,j=1}^N x_i^2 = 2\frac{1}{N}\sum_{i=1}^N x_i^2$,

since summation over j gives factor N.

The remaining term can be written as

$$2\frac{1}{N^2}\sum_{i,j=1}^N x_i x_j = 2\frac{1}{N^2}\sum_{i=1}^N x_i \sum_{i=1}^N x_j = 2\left(\frac{1}{N}\sum_{i=1}^N x_i\right)^2,$$

where we again used the fact that the sum is invariant to the index of summation.

We thus have proved the required relation that

$$\frac{1}{N^2}\sum_{i,j=1}^N (x_i - x_j)^2 = 2\frac{1}{N}\sum_{i=1}^N x_i^2 - 2\left(\frac{1}{N}\sum_{i=1}^N x_i\right)^2$$

Exercise 6.8

Bernoulli distribution is given by

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}$$

We can use relation

$$a^x = e^{x\log a}$$

to write the Bernoulli distribution as

$$p(x|\mu) = e^{x\log\mu + (1-x)\log(1-\mu)} = e^{x\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)} = h(x)e^{\theta x - A(\theta)},$$

where the last equation is the definition of a single-parameter distribution from the exponential family, in which

$h(x) = 1,$
$\theta = \log\left(\frac{\mu}{1-\mu}\right) \leftrightarrow \mu = \frac{e^\theta}{1+e^\theta},$
$A(\theta) = -\log(1-\mu) = \log(1+e^\theta)$

Exercise 6.9

The binomial distribution can be transformed as

$$p(x|N,\mu) = \binom{N}{x}\mu^x(1-\mu)^{N-x} = \binom{N}{x}e^{x\log\mu+(N-x)\log(1-\mu)} = \binom{N}{x}e^{x\log\left(\frac{\mu}{1-\mu}\right)+N\log(1-\mu)} = h(x)e^{x\theta-A(\theta)}$$

where

$$h(x) = \binom{N}{x},$$
$$\theta = \log\left(\frac{\mu}{1-\mu}\right),$$
$$A(\theta) = -N\log(1-\mu) = N\log(1+e^\theta)$$

i.e., the binomial distribution can be represented as an exponential family distribution(only μ is treated here as a parameter, since the number of trials N is fixed.)

Similarly, the beta distribution can be transoformed as

$$p(x|\alpha,\beta) = \frac{1}{\mathcal{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} = e^{(\alpha-1)\log x+(\beta-1)\log(1-x)-\log(\mathcal{B}(\alpha,\beta))} = h(x)e^{\theta_1\phi_1(x)+\theta_2\phi_2(x)-A(\theta_1,\theta_2)}$$

where

$$h(x) = 1,$$
$$\theta_1 = \alpha-1, \theta_2 = \beta-1,$$
$$\phi_1(x) = \log x, \phi_2(x) = \log(1-x),$$
$$A(\theta_1,\theta_2) = \log(\mathcal{B}(\alpha,\beta)) = \log(\mathcal{B}(1+\theta_1, 1+\theta_2))$$

i.e. this is a distribution form the exponential family.

The product of the two distributions is then given by

$$p(x|N,\mu)p(x|\alpha,\beta) = \binom{N}{x}e^{x\log\left(\frac{\mu}{1-\mu}\right)+(\alpha-1)\log x+(\beta-1)\log(1-x)+N\log(1-\mu)-\log(\mathcal{B}(\alpha,\beta))} = h(x)e^{\theta_1\phi_1(x)+\theta_2\phi_2(x)+\theta_3\phi_3(x$$

where

$$h(x) = \binom{N}{x},$$
$$\theta_1 = \alpha-1, \theta_2 = \beta-1, \theta_3 = \log\left(\frac{\mu}{1-\mu}\right)$$
$$\phi_1(x) = \log x, \phi_2(x) = \log(1-x), \phi_3(x) = x$$
$$A(\theta_1,\theta_2,\theta_3) = \log(\mathcal{B}(\alpha,\beta)) - N\log(1-\mu) = \log(\mathcal{B}(1+\theta_1, 1+\theta_2)) + N\log(1+e_3^\theta)$$

Exercise 6.10

a) ¶

The two normal distributions are given by

$$\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A}) = (2\pi)^{-\frac{D}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a})\right],$$
$$\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) = (2\pi)^{-\frac{D}{2}}|\mathbf{B}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b})\right]$$

their product is

$$\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) = (2\pi)^{-D}|\mathbf{AB}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}\left[(\mathbf{x}-\mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b})\right]\Big\}$$

The expression in the exponent can be written as

$$\Phi = (\mathbf{x}-\mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b}) =$$
$$\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} - \mathbf{a}^T\mathbf{A}^{-1}\mathbf{x} - \mathbf{x}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{x}^T\mathbf{B}^{-1}\mathbf{x} - \mathbf{b}^T\mathbf{B}^{-1}\mathbf{x} - \mathbf{x}^T\mathbf{B}^{-1}\mathbf{b} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} =$$
$$\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x} - (\mathbf{a}^T\mathbf{A}^{-1}+\mathbf{b}^T\mathbf{B}^{-1})\mathbf{x} - \mathbf{x}^T(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) + \mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}$$

we now introduce notation

$$\mathbf{C}^{-1} = (\mathbf{A}^{-1}+\mathbf{B}^{-1}),$$
$$\mathbf{c} = \mathbf{C}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}),$$
$$\mathbf{c}^T = (\mathbf{a}^T\mathbf{A}^{-1}+\mathbf{b}^T\mathbf{B}^{-1})C \text{ (This can be checked by transposing the previous equation)}$$

The expression in the exponent now takes form

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

where we have completed the square.

The product of the two probability distributions can be now written as

$$\mathcal{N}(\mathbf{x}|\mathbf{a},\mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b},\mathbf{B}) = (2\pi)^{-D}|\mathbf{AB}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}\left[(\mathbf{x}-\mathbf{c})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c}) + \mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{C}^{-1}\mathbf{c}\right]\Big\} =$$

$$(2\pi)^{-\frac{D}{2}}|\mathbf{C}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{c})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c})\right] \times (2\pi)^{-\frac{D}{2}}\frac{|\mathbf{AB}|^{-\frac{1}{2}}}{|\mathbf{C}|^{-\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\left[\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{C}^{-1}\mathbf{c}\right]\Big\} =$$

$$c\mathcal{N}(\mathbf{c}|\mathbf{c},\mathbf{C}),$$

where we defined

$$c = (2\pi)^{-\frac{D}{2}}\frac{|\mathbf{AB}|^{-\frac{1}{2}}}{|\mathbf{C}|^{-\frac{1}{2}}}\exp\Big\{-\frac{1}{2}\left[\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{C}^{-1}\mathbf{c}\right]\Big\}$$

We now can used the properties that a) the determinant of a matrix product is product of the determinants, and b) determinant of a matrix inverse is the inverse of the determinant of this matrix, and write

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

For the expression in the exponent we can write

$$\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{C}^{-1}\mathbf{c} =$$
$$\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - (\mathbf{a}^T\mathbf{A}^{-1}+\mathbf{b}^T\mathbf{B}^{-1})(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}) =$$
$$\mathbf{a}^T\left[\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}^{-1}\right]\mathbf{a} + \mathbf{b}^T\left[\mathbf{B}^{-1} - \mathbf{B}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{B}^{-1}\right]\mathbf{b} - \mathbf{a}^T\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}\mathbf{b} -$$

Using the property $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ we obtain

$$\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} = \left[\mathbf{B}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{A}\right]^{-1} = (\mathbf{A}+\mathbf{B})^{-1}$$

and

$$\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})\mathbf{A}^{-1} = \mathbf{A}^{-1}\left[1 - (\mathbf{A}^{-1} + \mathbf{B}^{-1})\mathbf{A}^{-1}\right] = \mathbf{A}^{-1}\left[1 - \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}\mathbf{A}^{-1}\right] = \mathbf{A}^{-1}\left[1 - \mathbf{B}(\right.$$
$$= \mathbf{A}^{-1}\left[(\mathbf{A} + \mathbf{B}) - \mathbf{B}\right](\mathbf{A} + \mathbf{B})^{-1} = (\mathbf{A} + \mathbf{B})^{-1}$$

we thus conclude that

$$c = (2\pi)^{-\frac{D}{2}}|\mathbf{A} + \mathbf{B}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{a} - \mathbf{b})^T(\mathbf{A} + \mathbf{B})^{-1}(\mathbf{a} - \mathbf{b})\right\} = \mathcal{N}(\mathbf{b}|\mathbf{a}, \mathbf{A} + \mathbf{B}) = \mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{A} + \mathbf{B})$$

b)

Multivariate normal distribution, $\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})$ can be represented as a distribution from an exponential family:

$$\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A}) = (2\pi)^{-\frac{D}{2}}|\mathbf{A}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x} - \mathbf{a})\right] =$$
$$(2\pi)^{-\frac{D}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}(\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T) + \mathbf{a}^T\mathbf{A}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\log|\mathbf{A}|\right],$$

where we used that $\mathbf{a}^T\mathbf{A}^{-1}\mathbf{x} = \mathbf{x}^T\mathbf{A}^{-1}\mathbf{a}$, and also write the first term as

$$\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} = \sum_{i,j} x_i(\mathbf{A}^{-1})_{ij}x_j = \sum_{i,j}(\mathbf{A}^{-1})_{ij}x_j x_i = \sum_{i,j}(\mathbf{A}^{-1})_{ij}(\mathbf{x}\mathbf{x}^T)_{ji} = \mathrm{tr}(\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^T)$$

Representing $\mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{B})$ in a similar way and multiplying the two distributions we readily obtain

$$\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{B}) = (2\pi)^{-D}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{A}^{-1} + \mathbf{B}^{-1})\mathbf{x}\mathbf{x}^T\right] + (\mathbf{a}^T\mathbf{A}^{-1} + \mathbf{b}^T\mathbf{B}^{-1})\mathbf{x} - \frac{1}{2}\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\log|\right.$$
$$c\mathcal{N}(\mathbf{x}|\mathbf{c}, \mathbf{C}),$$

where we defined

$$\mathbf{C}^{-1} = \mathbf{A}^{-1} + \mathbf{B}^{-1},$$
$$\mathbf{c}^T\mathbf{C}^{-1} = \mathbf{a}^T\mathbf{A}^{-1} + \mathbf{b}^T\mathbf{B}^{-1},$$
$$c = (2\pi)^{-\frac{D}{2}}\exp\left\{\frac{1}{2}\mathbf{c}^T\mathbf{C}^{-1}\mathbf{c} + \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} - \frac{1}{2}\log|\mathbf{A}| - \frac{1}{2}\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} - \frac{1}{2}\log|\mathbf{B}|\right\}$$

Coefficient c can now be reduced to the required form using the matrix transformations described in part a).

Exercise 6.11

The expectation value and the conditional expectation value are given by

$$\mathbb{E}_X[x] = \int xp(x)dx,$$
$$\mathbb{E}_Y[f(y)] = \int f(y)p(y)dy,$$
$$\mathbb{E}_X[x|y] = \int xp(x|y)dx$$

We then have

$$\mathbb{E}_Y\left[\mathbb{E}_X[x|y]\right] = \int \mathbb{E}_X[x|y]p(y)dy = \int\left[\int xp(x|y)dx\right]p(y)dy = \int\int xp(x|y)p(y)dxdy = \int\int xp(x,y)dxd$$
$$= \mathbb{E}_X[x],$$

where we used the definition fo the conditional probability density

$$p(x|y)p(y) = p(x, y)$$

Exercise 6.12

a)

If $\mathbf{x}$ is fixed, then $\mathbf{y}$ has the same distribution as $\mathbf{w}$, but with the mean shifter by $\mathbf{Ax} + \mathbf{b}$, that is

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{Q})$$

b)

Let us consider random variable $\mathbf{u} = \mathbf{Ax}$, it is distributed according to

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{A}\mu_x, \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T).$$

Then $\mathbf{y}$ is a sum of two Gaussian random variables $\mathbf{u}$ and $\mathbf{w}$ with its mean additionally shifted by $\mathbf{b}$, that is

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mu_x + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T + \mathbf{Q}),$$

that is

$$\mu_y = \mathbf{A}\mu_x + \mathbf{b},$$
$$\mathbf{\Sigma}_y = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T + \mathbf{Q}.$$

c)

Like in b), assuming that $\mathbf{y}$ is fixed we obtain the conditional distribution

$$p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\mathbf{Cy}, \mathbf{R})$$

Since $\mathbf{Cy}$ is a Gausssian random variable with distribution $\mathcal{N}(\mathbf{C}\mu_y, \mathbf{C}\mathbf{\Sigma}_y\mathbf{C}^T)$ we obtain the distribution of $\mathbf{z}$ as that of a sum of two Gaussian random variables:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{C}\mu_y, \mathbf{C}\mathbf{\Sigma}_y\mathbf{C}^T + \mathbf{R}) = \mathcal{N}(\mathbf{z}|\mathbf{C}(\mathbf{A}\mu_x + \mathbf{b}), \mathbf{C}(\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T + \mathbf{Q})\mathbf{C}^T + \mathbf{R})$$

d)

The posterior distribution $p(\mathbf{x}|\mathbf{y})$ can be obtained by applying the Bayes' theorem:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{\mathcal{N}(\mathbf{y}|\mathbf{Ax}+\mathbf{b},\mathbf{Q})\mathcal{N}(\mathbf{x}|\mu_x,\mathbf{\Sigma}_x)}{\mathcal{N}(\mathbf{y}|\mathbf{A}\mu_x+\mathbf{b},\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^T+\mathbf{Q})}$$

Exercise 6.13

Cdf is related to pdf as

$$F_x(x) = \int_{-\infty}^{x} dx' f_x(x'),$$
$$\frac{d}{dx}F_x(x) = f_x(x)$$

and changes in the interval $[0, 1]$.

The pdf of variable $y = F_x(x)$ then can be defined as

$$f_y(y) = f_x(x)\left|\frac{dx}{dy}\right| = \frac{f_x(x)}{\left|\frac{dy}{dx}\right|} = \frac{f_x(x)}{\left|\frac{dFx(x)}{dx}\right|} = \frac{f_x(x)}{f_x(x)} = 1,$$

i.e. y is uniformly distributed in interval $[0, 1]$.