Probabilistic Modeling and Reasoning

Homework — 3

Nikolaos Liouliakis (AID25001)
Vasileios-Efraim Tsavalia (AID25006)

MSc in Artificial Intelligence and Data Analytics
University of Macedonia

Supervisor: Professor Dimitris Christou-Varsakelis

February 2025

Problem 1

Given:

$$p(x,y) \propto (x^2+y^2)^2 e^{-x^2-y^2} \quad dom(x) = dom(y) = \{-\inf, \ldots, inf\}$$

Show that:

$$\langle x \rangle = \langle y \rangle = 0$$

$$\langle xy \rangle = \langle x \rangle \langle y \rangle$$

$$p(x,y) \neq p(x)p(y)$$

Find c such that $p(x,y)$ is a distribution:

$$p(x,y) = \frac{1}{c}(x^2+y^2)^2 e^{-x^2-y^2}$$

$$c = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x^2+y^2)^2 e^{-x^2-y^2}\, dxdy = 2\pi$$

Therefore

$$p(x,y) = \frac{1}{2\pi}(x^2+y^2)^2 e^{-x^2-y^2} \tag{1}$$

We know that

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2}\, dx = \sqrt{\frac{\pi}{a}} \tag{2}$$

Using the above and by differentiating under the integral sign, it is apparent that:

$$\begin{aligned}
\int_{-\infty}^{\infty} x^{2n} e^{-\alpha x^2}\, dx &= (-1)^n \int_{-\infty}^{\infty} \frac{\partial^n}{\partial \alpha^n} e^{-\alpha x^2}\, dx \\
&= (-1)^n \frac{\partial^n}{\partial \alpha^n} \int_{-\infty}^{\infty} e^{-\alpha x^2}\, dx \\
&= \sqrt{\pi}\, (-1)^n \frac{\partial^n}{\partial \alpha^n} \alpha^{-\frac{1}{2}} \\
&= \sqrt{\frac{\pi}{\alpha}} \frac{(2n-1)!!}{(2\alpha)^n}
\end{aligned} \tag{3}$$

$$
\begin{aligned}
p(x) &= \int_{-\infty}^{\infty} p(x,y)\,dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi}(x^2+y^2)^2 e^{-x^2-y^2}\,dy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} (x^4+2x^2y^2+y^4)e^{-x^2-y^2}\,dy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} x^4 e^{-x^2-y^2}\,dy + \frac{1}{2\pi}\int_{-\infty}^{\infty} 2x^2y^2 e^{-x^2-y^2}\,dy + \frac{1}{2\pi}\int_{-\infty}^{\infty} y^4 e^{-x^2-y^2}\,dy \\
&= \frac{1}{2\pi}x^4 e^{-x^2}\int_{-\infty}^{\infty} e^{-y^2}\,dy + \frac{1}{2\pi}2x^2 e^{-x^2}\int_{-\infty}^{\infty} y^2 e^{-y^2}\,dy + \frac{1}{2\pi}e^{-x^2}\int_{-\infty}^{\infty} y^4 e^{-y^2}\,dy \\
&= \frac{1}{2\pi}x^4 e^{-x^2}\sqrt{\pi} + \frac{1}{2\pi}2x^2 e^{-x^2}\frac{\sqrt{\pi}}{2} + \frac{1}{2\pi}e^{-x^2}\frac{3\sqrt{\pi}}{4} \\
&= \frac{\sqrt{\pi}}{2\pi}(x^4+x^2+\frac{3}{4})e^{-x^2}
\end{aligned}
$$

Because $p(x)$ is an even function, $xp(x)$ is an odd function, therefore it follows that:

$$\langle x \rangle = \int_{-\infty}^{\infty} xp(x)\,dx$$

$$\langle x \rangle = \int_{-\infty}^{\infty} x\frac{\sqrt{\pi}}{2\pi}(x^4+x^2+\frac{3}{4})e^{-x^2}\,dx = \int_{-\infty}^{\infty} \frac{\sqrt{\pi}}{2\pi}(x^5+x^3+\frac{3}{4}x)e^{-x^2}\,dx = 0$$

Similarly, we can calculate $p(y)$ and $\langle y \rangle$ as:

$$p(y) = \frac{\sqrt{\pi}}{2\pi}(y^4+y^2+\frac{3}{4})e^{-y^2}$$

$$\langle y \rangle = 0$$

To check if they are uncorrelated, we must find the covariance. To do so, we first find $\langle xy \rangle$

$$\langle xy \rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)\,dx\,dy$$

$$\langle xy \rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\frac{1}{2\pi}(x^2+y^2)^2 e^{-x^2-y^2}\,dx\,dy = 0$$

Which holds because $xyp(x,y)$ is odd with respect to both x and y.

Thus, x and y are uncorrelated because:

$$cov(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle = 0$$

But they are dependent because:

$$
\begin{aligned}
p(x,y) &\neq p(x)p(y) \\
\frac{1}{2\pi}(x^2+y^2)^2 e^{-x^2-y^2} &\neq \frac{\sqrt{\pi}}{2\pi}(x^4+x^2+\frac{3}{4})e^{-x^2}\frac{\sqrt{\pi}}{2\pi}(y^4+y^2+\frac{3}{4})e^{-y^2} \\
\frac{1}{2\pi}(x^2+y^2)^2 e^{-x^2-y^2} &\neq \frac{1}{4\pi}(x^4+x^2+\frac{3}{4})(y^4+y^2+\frac{3}{4})e^{-x^2-y^2}
\end{aligned}
$$

Problem 2

Part 1

Given: A random variable X with probability density function $f_X(x)$ and a strictly monotonic function $g : \mathbb{R} \to \mathbb{R}$.

Goal: Find the probability density function $f_Y(y)$ of $Y = g(X)$.

Since g is monotonic, we consider two cases: g is strictly increasing or strictly decreasing.

The cumulative distribution function (CDF) of Y is:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

Case 1: g is strictly increasing:
If g is strictly increasing, then $g(x) \leq y$ implies $x \leq g^{-1}(y)$. Thus,

$$F_Y(y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Case 2: g is strictly decreasing:
If g is strictly decreasing, then $g(x) \leq y$ implies $x \geq g^{-1}(y)$. Thus,

$$F_Y(y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

Differentiate to Find $f_Y(y)$:
In both cases, we differentiate $F_Y(y)$ with respect to y to find $f_Y(y)$:

- For an increasing g:

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \cdot \frac{d}{dy}\left(g^{-1}(y)\right)$$

- For a decreasing g:

$$f_Y(y) = \frac{d}{dy}\left(1 - F_X(g^{-1}(y))\right) = f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right|$$

By the inverse function theorem, $\frac{d}{dy} g^{-1}(y) = \frac{1}{g'(g^{-1}(y))}$. Therefore, in both cases we have:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{1}{g'(g^{-1}(y))}\right|$$

Thus, the probability density function $f_Y(y)$ is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|g'(g^{-1}(y))\right|^{-1}$$

Part 2

Let X be a random variable in $\mathbb{R}^p$ with joint probability density function $f_X(x)$ (with respect to the Lebesgue measure), so that for any (measurable) set $A \subset \mathbb{R}^p$,

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, dx.$$

Let $g : \mathbb{R}^p \to \mathbb{R}^p$ be a bijective, continuously differentiable function. We define a new random variable $Y = g(X)$ and aim to determine the probability density function $f_Y(y)$ of Y (with respect to the Lebesgue measure).

Step 1: Transformation and (Preimage Sets)

For any (measurable) set $B \subset \mathbb{R}^p$, we express the probability that Y takes a value in B in terms of X:

$$\mathbb{P}(Y \in B) = \mathbb{P}(g(X) \in B).$$

Since g is bijective, the event $g(X) \in B$ is equivalent to $X \in g^{-1}(B)$, where $g^{-1}(B)$ denotes the preimage of B under g. Therefore,

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in g^{-1}(B)) = \int_{g^{-1}(B)} f_X(x)\, dx.$$

Step 2: Change of Variables Using the Jacobian

The set $g^{-1}(B)$ is the preimage of B under g, and by the change of variables theorem in multivariable calculus, we can transform the integral over $g^{-1}(B)$ to an integral over B by using the Jacobian determinant of g.

Define the Jacobian matrix $J(x)$ of g at x by

$$J(x) = \left(\frac{\partial g_i}{\partial x_j}\right)_{i,j=1}^{p},$$

and let $\det J(x)$ denote its determinant. By the change of variables formula, we have

$$\int_{g^{-1}(B)} f_X(x)\, dx = \int_B f_X(g^{-1}(y)) \left|\det J(g^{-1}(y))\right|^{-1}\, dy.$$

Therefore,

$$\mathbb{P}(Y \in B) = \int_B f_X(g^{-1}(y)) \left|\det J(g^{-1}(y))\right|^{-1}\, dy.$$

Step 3: Identifying the Density $f_Y(y)$

The probability $\mathbb{P}(Y \in B)$ for any (measurable) set B can also be expressed in terms of the pdf $f_Y(y)$ of Y as

$$\mathbb{P}(Y \in B) = \int_B f_Y(y)\, dy.$$

By comparing this expression with the result from Step 2, we find that

$$f_Y(y) = f_X(g^{-1}(y)) \left|\det J(g^{-1}(y))\right|^{-1}.$$

Conclusion

We have derived that the pdf $f_Y(y)$ of the transformed random variable $Y = g(X)$ is given by

$$f_Y(y) = \frac{f_X(x)}{|\det J(x)|} \quad \text{where } x = g^{-1}(y).$$

In other words,

$$f_Y(y) = f_X(g^{-1}(y)) \left|\det J(g^{-1}(y))\right|^{-1}$$

Problem 3

To derive the differential entropy of a multivariate normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$, we start from the general definition of differential entropy.

1. Differential Entropy Definition

For a continuous random variable X with pdf $f(x)$, the differential entropy $h(X)$ is defined as:

$$h(X) = -\int_{\mathbb{R}^d} f(x) \log f(x)\, dx$$

2. Multivariate Normal Distribution PDF

For a d-dimensional multivariate normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$, the probability density function is:

$$f(x) = \frac{1}{(2\pi)^{d/2}\, |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

3. Substitute the PDF into the Entropy Formula

Substituting $f(x)$ into the entropy formula:

$$h(X) = -\int_{\mathbb{R}^d} f(x) \log f(x)\, dx$$

we get:

$$h(X) = -\int_{\mathbb{R}^d} \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{d/2}|\Sigma|^{1/2}} \log\left(\frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{d/2}|\Sigma|^{1/2}}\right) dx$$

4. Simplify the Logarithmic Term

Expanding the logarithm:

$$\log f(x) = \log\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\right) + \log\left(\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)\right)$$

yields:

$$\log f(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$$

5. Substitute and Separate Terms

Substitute $\log f(x)$ back into the entropy integral:

$$h(X) = -\int_{\mathbb{R}^d} f(x)\left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx$$

Distribute $f(x)$ and separate terms:

$$h(X) = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\int_{\mathbb{R}^d} f(x)(x-\mu)^\top \Sigma^{-1}(x-\mu)\, dx$$

6. Evaluate Integral Terms

Constant Terms
The first two terms are constants, unaffected by the integration and because $f(x)$ is a pdf they are simplified using:

$$\int_{\mathbb{R}^d} f(x)\, dx = 1$$

Quadratic Term
The last term represents the expected value of $(x-\mu)^\top \Sigma^{-1}(x-\mu)$ under $f(x)$, (the Mahalanobis distance

squared). Since X is multivariate normal with covariance Σ, this expectation evaluates to the trace of the identity matrix in d-dimensions, which is d.

$$\int_{\mathbb{R}^d} f(x)(x-\mu)^\top \Sigma^{-1}(x-\mu)\,dx = d$$

It can be proven by viewing the integral as an expected value:

$$\left\langle (x-\mu)^\top \Sigma^{-1}(x-\mu) \right\rangle_{\mathcal{N}(\mu,\Sigma)}$$

Step 1: Expand $(x-\mu)^\top \Sigma^{-1}(x-\mu)$

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = x^\top \Sigma^{-1} x - x^\top \Sigma^{-1}\mu - \mu^\top \Sigma^{-1} x + \mu^\top \Sigma^{-1}\mu$$

Since they are scalars, $x^\top \Sigma^{-1}\mu = \mu^\top \Sigma^{-1} x$ and we can rewrite it as:

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = x^\top \Sigma^{-1} x - 2\mu^\top \Sigma^{-1} x + \mu^\top \Sigma^{-1}\mu$$

Step 2: Take the Expectation Over $\mathcal{N}(\mu,\Sigma)$

Now we take the expectation over each term separately:

$$\left\langle (x-\mu)^\top \Sigma^{-1}(x-\mu) \right\rangle_{\mathcal{N}(\mu,\Sigma)} = \mathbb{E}\left[x^\top \Sigma^{-1} x\right] - 2\mathbb{E}\left[\mu^\top \Sigma^{-1} x\right] + \mathbb{E}\left[\mu^\top \Sigma^{-1}\mu\right]$$

Term 1: $\mathbb{E}\left[x^\top \Sigma^{-1} x\right]$

Since $x^\top \Sigma^{-1} x = \mathrm{Tr}(x^\top \Sigma^{-1} x) = \mathrm{Tr}(\Sigma^{-1} x x^\top)$, we have:

$$\mathbb{E}\left[x^\top \Sigma^{-1} x\right] = \mathbb{E}\left[\mathrm{Tr}(\Sigma^{-1} x x^\top)\right] = \mathrm{Tr}(\Sigma^{-1}\mathbb{E}[x x^\top])$$

Using $\mathbb{E}[x x^\top] = \Sigma + \mu\mu^\top$, this becomes:

$$\mathbb{E}\left[x^\top \Sigma^{-1} x\right] = \mathrm{Tr}(\Sigma^{-1}(\Sigma + \mu\mu^\top)) = \mathrm{Tr}(\Sigma^{-1}\Sigma) + \mathrm{Tr}(\Sigma^{-1}\mu\mu^\top)$$

Since $\mathrm{Tr}(\Sigma^{-1}\Sigma) = \mathrm{Tr}(I) = d$, where d is the dimension, and $\mathrm{Tr}(\Sigma^{-1}\mu\mu^\top) = \mathrm{Tr}(\mu^\top \Sigma^{-1}\mu) = \mu^\top \Sigma^{-1}\mu$, we get:

$$\mathbb{E}\left[x^\top \Sigma^{-1} x\right] = d + \mu^\top \Sigma^{-1}\mu$$

Term 2: $-2\mathbb{E}\left[\mu^\top \Sigma^{-1} x\right]$

Since $\mu^\top \Sigma^{-1}$ is constant with respect to x, we have:

$$-2\mathbb{E}\left[\mu^\top \Sigma^{-1} x\right] = -2\mu^\top \Sigma^{-1}\mathbb{E}[x] = -2\mu^\top \Sigma^{-1}\mu$$

Term 3: $\mathbb{E}\left[\mu^\top \Sigma^{-1}\mu\right]$

This term is constant, so its expectation is simply:

$$\mathbb{E}\left[\mu^\top \Sigma^{-1}\mu\right] = \mu^\top \Sigma^{-1}\mu$$

Step 3: Combine All Terms

Now we combine these results:

$$\left\langle (x-\mu)^\top \Sigma^{-1}(x-\mu) \right\rangle_{\mathcal{N}(\mu,\Sigma)} = d + \mu^\top \Sigma^{-1}\mu - 2\mu^\top \Sigma^{-1}\mu + \mu^\top \Sigma^{-1}\mu$$

Simplifying terms involving $\mu^\top \Sigma^{-1} \mu$, we get:

$$\left\langle (x-\mu)^\top \Sigma^{-1} (x-\mu) \right\rangle_{\mathcal{N}(\mu,\Sigma)} = d$$

Final Result

Thus, the differential entropy of a d-dimensional multivariate normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$ is:

$$h(X) = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\cdot d$$

or equivalently:

$$h(X) = \frac{1}{2}\log\left((2\pi e)^d |\Sigma|\right)$$

Problem 4

Let y be linearly related to x through $y = \mathbf{M}x + \eta$ where $x \perp \eta$, $\eta \sim \mathcal{N}(\mu, \Sigma)$, and $x \sim \mathcal{N}(\mu_x, \Sigma_x)$.

Prove that the marginal $p(y) = \int_x p(y|x)p(x)$ is a Gaussian:

$$p(y) = \mathcal{N}\left(y \mid \mathbf{M}\mu_x + \mu,\, \mathbf{M}\Sigma_x\mathbf{M}^\top + \Sigma\right)$$

Part 1

We start with the expression for $p(y)$:

$$p(y) = \int p(y|x)p(x)\,dx$$

$$p(y) \propto \int \exp\left(-\frac{1}{2}(y-Mx-\mu)^\top\Sigma^{-1}(y-Mx-\mu) - \frac{1}{2}(x-\mu_x)^\top\Sigma_x^{-1}(x-\mu_x)\right)dx$$

Expanding the Quadratic Forms

$$(y-Mx-\mu)^\top\Sigma^{-1}(y-Mx-\mu) = y^\top\Sigma^{-1}y - 2y^\top\Sigma^{-1}Mx + x^\top M^\top\Sigma^{-1}Mx - 2y^\top\Sigma^{-1}\mu + 2x^\top M^\top\Sigma^{-1}\mu + \mu^\top\Sigma^{-1}\mu$$

$$(x-\mu_x)^\top\Sigma_x^{-1}(x-\mu_x) = x^\top\Sigma_x^{-1}x - 2x^\top\Sigma_x^{-1}\mu_x + \mu_x^\top\Sigma_x^{-1}\mu_x$$

Combine Terms

- Terms involving only y:

$$y^\top\Sigma^{-1}y - 2y^\top\Sigma^{-1}\mu$$

- Quadratic terms in x:

$$x^\top(M^\top\Sigma^{-1}M + \Sigma_x^{-1})x$$

- Linear terms in x:

$$x^\top\left(-2M^\top\Sigma^{-1}y + 2M^\top\Sigma^{-1}\mu + 2\Sigma_x^{-1}\mu_x\right)$$

- Constant terms :

$$\mu^\top\Sigma^{-1}\mu + \mu_x^\top\Sigma_x^{-1}\mu_x$$

The terms involving only y can be factored outside the integral:

$$p(y) \propto \exp\left(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu\right)\int \exp\left(-\frac{1}{2}x^\top Ax + x^\top(By+c)\right)dx$$

where we define:

$$\begin{aligned} A &= M^\top\Sigma^{-1}M + \Sigma_x^{-1} \\ B &= M^\top\Sigma^{-1} \\ c &= \Sigma_x^{-1}\mu_x - M^\top\Sigma^{-1}\mu \end{aligned}$$

Part 2

Because $x^\top(By+c)$ is scalar, it is equal with his transpose $(By+c)^\top x$ therefore:

$$p(y) \propto \exp\left(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu\right)\int \exp\left(-\frac{1}{2}x^\top Ax + (By+c)^\top x\right)dx$$

Using the identity

$$\int \exp\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{x}\right)\,d\mathbf{x} = \sqrt{\det(2\pi\mathbf{A}^{-1})}\exp\left(\frac{1}{2}\mathbf{b}^\top\mathbf{A}^{-1}\mathbf{b}\right)$$

We can compute the integral:

$$\int \exp\left(-\frac{1}{2}x^\top Ax + (By+c)^\top x\right)dx = \sqrt{\det(2\pi A^{-1})}\exp\left(\frac{1}{2}(By+c)^\top A^{-1}(By+c)\right)$$

Then we substitute to simplify the expression for $p(y)$

$$p(y) \propto \exp\left(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu\right)\sqrt{\det(2\pi A^{-1})}\exp\left(\frac{1}{2}(By+c)^\top A^{-1}(By+c)\right)$$

$$p(y) \propto \exp\left(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu\right)\exp\left(\frac{1}{2}(By+c)^\top A^{-1}(By+c)\right)$$

$$p(y) \propto \exp\left(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu + \frac{1}{2}(By+c)^\top A^{-1}(By+c)\right)$$

$$\begin{aligned} p(y) &\propto \exp\Bigg(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu + \frac{1}{2}(By+c)^\top A^{-1}(By+c)\Bigg) \\ &= \exp\Bigg(-\frac{1}{2}y^\top\Sigma^{-1}y + y^\top\Sigma^{-1}\mu + \frac{1}{2}(y^\top B^\top A^{-1}By + 2y^\top B^\top A^{-1}c + c^\top A^{-1}c)\Bigg) \\ &= \exp\Bigg(-\frac{1}{2}y^\top(\Sigma^{-1} - B^\top A^{-1}B)y + y^\top(\Sigma^{-1}\mu + B^\top A^{-1}c) + \frac{1}{2}c^\top A^{-1}c\Bigg) \end{aligned}$$

Therefore, Y follows a Gaussian distribution.

Part 3

The mean of Y is:

$$\langle y \rangle = \mathbf{M}\langle x \rangle + \langle \eta \rangle = \mathbf{M}\mu_x + \mu$$

The covariance of Y is:

$$\begin{aligned}
\left\langle (y - \langle y \rangle)\,(y - \langle y \rangle)^\top \right\rangle &= \left\langle (\mathbf{M}x + \eta - \mathbf{M}\mu_x - \mu)\,(\mathbf{M}x + \eta - \mathbf{M}\mu_x - \mu)^\top \right\rangle \\
&= \left\langle (\mathbf{M}(x-\mu_x) + (\eta - \mu))\,(\mathbf{M}(x-\mu_x) + (\eta - \mu))^\top \right\rangle \\
&= \langle \mathbf{M}(x-\mu_x)(x-\mu_x)^\top \mathbf{M}^\top + \mathbf{M}(x-\mu_x)(\eta-\mu)^\top \\
&\quad + (\eta-\mu)(x-\mu_x)^\top \mathbf{M}^\top + (\eta-\mu)(\eta-\mu)^\top \rangle \\
&= \mathbf{M}\langle (x-\mu_x)(x-\mu_x)^\top \rangle \mathbf{M}^\top + \mathbf{M}\langle (x-\mu_x)(\eta-\mu)^\top \rangle \\
&\quad + \langle (\eta-\mu)(x-\mu_x)^\top \rangle \mathbf{M}^\top + \langle (\eta-\mu)(\eta-\mu)^\top \rangle
\end{aligned}$$

The cross-covariance terms $\langle (x-\mu_x)(\eta-\mu)^\top \rangle$ and $\langle (\eta-\mu)(x-\mu_x)^\top \rangle$ are zero since x and η are independent.

$$\begin{aligned}
\left\langle (y - \langle y \rangle)\,(y - \langle y \rangle)^\top \right\rangle &= \mathbf{M}\langle (x-\mu_x)(x-\mu_x)^\top \rangle \mathbf{M}^\top + \mathbf{M}\langle (x-\mu_x)(\eta-\mu)^\top \rangle \\
&\quad + \langle (\eta-\mu)(x-\mu_x)^\top \rangle \mathbf{M}^\top + \langle (\eta-\mu)(\eta-\mu)^\top \rangle \\
&= \mathbf{M}\langle (x-\mu_x)(x-\mu_x)^\top \rangle \mathbf{M}^\top + \langle (\eta-\mu)(\eta-\mu)^\top \rangle \\
&= \mathbf{M}\Sigma_x\mathbf{M}^\top + \Sigma,
\end{aligned}$$

Therefore:

$$p(y) = \mathcal{N}\left(y \mid \mathbf{M}\mu_x + \mu,\, \mathbf{M}\Sigma_x\mathbf{M}^\top + \Sigma\right)$$

Problem 5

To find the maximum likelihood estimator (MLE) of the Poisson parameter λ, we start by writing the probability mass function of the Poisson distribution:

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Given a sample $x_1, x_2, \ldots, x_n$ drawn independently from this distribution, the likelihood function $p(\lambda \mid \boldsymbol{x})$ is the joint probability of observing this sample:

$$p(\lambda \mid \boldsymbol{x}) \propto \prod_{i=1}^{n} p(x_i \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$p(\lambda \mid \boldsymbol{x}) \propto \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

To find the MLE of λ (which maximizes $p(\lambda \mid \boldsymbol{x})$), we take the derivative of $p(\lambda \mid \boldsymbol{x})$ with respect to λ and set it to zero:

$$\frac{p(\lambda \mid \boldsymbol{x})}{d\lambda} \propto e^{-n\lambda}\left((\sum_{i=1}^{n} x_i)\lambda^{(\sum_{i=1}^n x_i)-1} - n\lambda^{\sum_{i=1}^n x_i} \right) = 0$$

$$\lambda^{(\sum_{i=1}^{n} x_i)-1}\left(\sum_{i=1}^{n} x_i - n\lambda\right) = 0$$

Solving for λ, we get:

$$\lambda = 0 \quad \text{or} \quad \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It is apparant that $\lambda = 0$ is a minimum, thus we check the second derivative for $\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\left.\frac{d^2 p(\lambda \mid \boldsymbol{x})}{d\lambda^2}\right|_{\lambda=\frac{1}{n}\sum_{i=1}^{n} x_i} < 0$$

Therefore, the maximum likelihood estimator of λ is:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

Problem 6

Given this Gaussian mixture model

$$p(\mathbf{x}) = \sum_i p_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad p_i > 0, \quad \sum_i p_i = 1$$

The mean is:

$$\begin{aligned} \boldsymbol{\mu}_{\text{GMM}} &= \mathbb{E}[\mathbf{x}] = \mathbb{E}[\sum_i p_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)] \\ &= \sum_i p_i \mathbb{E}[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)] \\ &= \sum_i p_i \boldsymbol{\mu}_i \end{aligned} \tag{4}$$

The covariance can be calculated as:

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{GMM}} &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top] \\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top \end{aligned} \tag{5}$$

We know that:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top|i] = \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top$$

Therefore, we can calculate the second moment:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \sum_i p_i \mathbb{E}[\mathbf{x}\mathbf{x}^\top|i]$$

Substituting:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \sum_i p_i(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top)$$

Then, substituting the mean from (4):

$$\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top = \left(\sum_i p_i \boldsymbol{\mu}_i\right)\left(\sum_j p_j \boldsymbol{\mu}_j^\top\right)$$

Finally, combining the above results and substituting in (5), we get:

$$\boldsymbol{\Sigma}_{\text{GMM}} = \sum_i p_i(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top) - \left(\sum_i p_i \boldsymbol{\mu}_i\right)\left(\sum_j p_j \boldsymbol{\mu}_j^\top\right)$$