EECE7397 – Homework 2

1.

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \\ \mathbf{x}_c \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}, \quad \Sigma^{-1} = \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} & \Lambda_{ac} \\ \Lambda_{ba} & \Lambda_{bb} & \Lambda_{bc} \\ \Lambda_{ca} & \Lambda_{cb} & \Lambda_{cc} \end{pmatrix}.$$

$$\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \\ \mathbf{x}_c \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{pmatrix}, \Sigma \Big).$$

Since $\mathbf{x}_c$ is marginalized, $\Lambda_{ac}, \Lambda_{bc}, \Lambda_{ca}, \Lambda_{cb}, \Lambda_{cc}$ do not affect the marginal distribution in $\mathbf{x}_a$ and $\mathbf{x}_b$. Thus,

$$\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \Big).$$

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} \, |\Sigma|^{1/2}} \exp\Big\{ -\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \Big\}.$$

$$\begin{aligned}
&-\frac{1}{2} \big(\mathbf{x} - \boldsymbol{\mu}\big)^\top \Sigma^{-1} \big(\mathbf{x} - \boldsymbol{\mu}\big) \\
&= -\frac{1}{2} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix}^\top \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix} \\
&= -\tfrac{1}{2} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big)^\top \Lambda_{aa} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big) - \tfrac{1}{2} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big)^\top \Lambda_{ab} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big) \\
&\quad - \tfrac{1}{2} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big)^\top \Lambda_{ba} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big) - \tfrac{1}{2} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big)^\top \Lambda_{bb} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big) \\
&= -\tfrac{1}{2} \Big[(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \Lambda_{aa} (\mathbf{x}_a - \boldsymbol{\mu}_a) + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \Lambda_{bb} (\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&\qquad + (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \Lambda_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a) \Big] \\
&= -\tfrac{1}{2} \Big[(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \Lambda_{aa} (\mathbf{x}_a - \boldsymbol{\mu}_a) + (\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \Lambda_{bb} (\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&\qquad + 2 (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \Big] \\
&= -\tfrac{1}{2} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big)^\top \Lambda_{aa} \big(\mathbf{x}_a - \boldsymbol{\mu}_a\big) - \tfrac{1}{2} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big)^\top \Lambda_{bb} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big) \\
&\quad - (\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \Lambda_{ab} \big(\mathbf{x}_b - \boldsymbol{\mu}_b\big).
\end{aligned}$$

Let

$$\mathbf{a} = \mathbf{x}_a - \boldsymbol{\mu}_a \quad \text{and} \quad \mathbf{b} = \mathbf{x}_b - \boldsymbol{\mu}_b.$$

$$
\begin{aligned}
&-\frac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\Lambda_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right) \;-\; \frac{1}{2}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right)^{\top}\Lambda_{bb}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right) \\
&\quad -\; \left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\Lambda_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right) \\
&\quad = -\frac{1}{2}\,\mathbf{a}^{\top}\Lambda_{aa}\,\mathbf{a} \;-\; \frac{1}{2}\,\mathbf{b}^{\top}\Lambda_{bb}\,\mathbf{b} \;-\; \mathbf{a}^{\top}\Lambda_{ab}\,\mathbf{b} \\
&\quad = -\frac{1}{2}\,\mathbf{a}^{\top}\Lambda_{aa}\,\mathbf{a} \;-\; \mathbf{a}^{\top}\Lambda_{ab}\,\mathbf{b} \qquad \left(\tfrac{1}{2}\,\mathbf{b}^{\top}\Lambda_{bb}\,\mathbf{b} \text{ is independent of } \mathbf{a}\right) \\
&\quad = -\tfrac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\Lambda_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right) \;-\; \left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\Lambda_{ab}\,\mathbf{b} \\
&\quad = -\tfrac{1}{2}\,\mathbf{x}_a^{\top}\Lambda_{aa}\,\mathbf{x}_a \;+\; \mathbf{x}_a^{\top}\Lambda_{aa}\,\boldsymbol{\mu}_a \;-\; \tfrac{1}{2}\,\boldsymbol{\mu}_a^{\top}\Lambda_{aa}\,\boldsymbol{\mu}_a \;-\; \mathbf{x}_a^{\top}\Lambda_{ab}\,\mathbf{b} \;+\; \boldsymbol{\mu}_a^{\top}\Lambda_{ab}\,\mathbf{b} \\
&\quad = -\tfrac{1}{2}\,\mathbf{x}_a^{\top}\Lambda_{aa}\,\mathbf{x}_a \;+\; \mathbf{x}_a^{\top}\left[\Lambda_{aa}\,\boldsymbol{\mu}_a \;-\; \Lambda_{ab}\,\mathbf{b}\right] \;-\; \tfrac{1}{2}\,\boldsymbol{\mu}_a^{\top}\Lambda_{aa}\,\boldsymbol{\mu}_a \;+\; \boldsymbol{\mu}_a^{\top}\Lambda_{ab}\,\mathbf{b}. \longrightarrow (1)
\end{aligned}
$$

$$
-\tfrac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}\right)^{\top}\Lambda_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}\right) \;=\; -\tfrac{1}{2}\,\mathbf{x}_a^{\top}\Lambda_{aa}\,\mathbf{x}_a \;+\; \mathbf{x}_a^{\top}\Lambda_{aa}\,\boldsymbol{\mu} \;-\; \tfrac{1}{2}\,\boldsymbol{\mu}^{\top}\Lambda_{aa}\,\boldsymbol{\mu}. \longrightarrow (2)
$$

From (1) and (2), comparing the term $\mathbf{x}_a^{\top}\Lambda_{aa}\,\boldsymbol{\mu}$ with $\mathbf{x}_a^{\top}\left[\Lambda_{aa}\,\boldsymbol{\mu}_a-\Lambda_{ab}\,\mathbf{b}\right]$ leads to:

$$
\begin{aligned}
&\Lambda_{aa}\,\boldsymbol{\mu} \;=\; \Lambda_{aa}\,\boldsymbol{\mu}_a \;-\; \Lambda_{ab}\,\mathbf{b}, \\
&\boldsymbol{\mu} \;=\; \Lambda_{aa}^{-1}\left(\Lambda_{aa}\,\boldsymbol{\mu}_a-\Lambda_{ab}\,\mathbf{b}\right) \;=\; \boldsymbol{\mu}_a \;-\; \Lambda_{aa}^{-1}\,\Lambda_{ab}\,\mathbf{b}.
\end{aligned}
$$

Substituting $\mathbf{b} \;=\; \mathbf{x}_b-\boldsymbol{\mu}_b$ gives the conditional mean:

$$
\boldsymbol{\mu}_{a|b} \;=\; \boldsymbol{\mu}_a \;-\; \Lambda_{aa}^{-1}\,\Lambda_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right).
$$

Focusing on the quadratic term in $\mathbf{x}_a$:

$$
-\tfrac{1}{2}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)^{\top}\Lambda_{aa}\left(\mathbf{x}_a-\boldsymbol{\mu}_a\right)
$$

shows that Λ_{aa} is the precision w.r.t $\mathbf{x}_a$, thus

$$
\Sigma_{a|b}^{-1} \;=\; \Lambda_{aa} \quad\Longrightarrow\quad \Sigma_{a|b} \;=\; \Lambda_{aa}^{-1}.
$$

Therefore, the conditional distribution is

$$
p\big(\mathbf{x}_a \mid \mathbf{x}_b\big) \;=\; \mathcal{N}\Big(\boldsymbol{\mu}_a-\Lambda_{aa}^{-1}\,\Lambda_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right),\; \Lambda_{aa}^{-1}\Big).
$$

2.

If $(A+B\,C\,D)^{-1}=A^{-1}\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}$, then by the definition of the inverse $AA^{-1}x=x$, for every vector x it must be that $\left(A+B\,C\,D\right)\Big[A^{-1}\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}\Big]\,x\;=\;x$

$$
\text{Let } y=\Big[A^{-1}\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}\Big]\,x
$$

$$
\begin{aligned}
y &= \Big[A^{-1}\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}\Big]\,x \\
&= A^{-1}x\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}x
\end{aligned}
$$

$$
\begin{aligned}
&(A+B\,C\,D)\cdot y \\
&= (A+B\,C\,D)\cdot\Big[A^{-1}x\;-\;A^{-1}\,B\left(C^{-1}+D\,A^{-1}\,B\right)^{-1}D\,A^{-1}x\Big] \\
&= (A+B\,C\,D)\Big[A^{-1}x\Big]\;-\;(A+B\,C\,D)\Big[A^{-1}B\,(C^{-1}+D\,A^{-1}B)^{-1}\,D\,A^{-1}x\Big]
\end{aligned}
$$

$$
\begin{aligned}
&(A + B\,C\,D)\Big[A^{-1}x\Big] \\
&= A\,A^{-1}x + B\,C\,D\big[A^{-1}x\big] \\
&= x + BCDA^{-1}x
\end{aligned}
$$

$$
\begin{aligned}
&(A + B\,C\,D)\Big[A^{-1}B\,(C^{-1} + D\,A^{-1}B)^{-1}\,D\,A^{-1}x\Big] \\
&= \Big[A\,A^{-1}B \;+\; B\,C\,D\,A^{-1}B\Big]\big(C^{-1} + D\,A^{-1}B\big)^{-1}D\,A^{-1}x \\
&= \Big[B \;+\; B\,C\,D\,A^{-1}B\Big]\big(C^{-1} + D\,A^{-1}B\big)^{-1}D\,A^{-1}x \\
&= B\Big[1 \;+\; C\,D\,A^{-1}B\Big]\big(C^{-1} + D\,A^{-1}B\big)^{-1}D\,A^{-1}x \\
&= B\Big[CC^{-1} \;+\; C\,D\,A^{-1}B\Big]\big(C^{-1} + D\,A^{-1}B\big)^{-1}D\,A^{-1}x \\
&= B\Big[C(C^{-1} \;+\; D\,A^{-1}B)\Big]\big(C^{-1} + D\,A^{-1}B\big)^{-1}D\,A^{-1}x \\
&= BC\Big[(C^{-1} \;+\; D\,A^{-1}B)\big(C^{-1} + D\,A^{-1}B\big)^{-1}\Big]D\,A^{-1}x \\
&= BCDA^{-1}x
\end{aligned}
$$

$$
\begin{aligned}
&(A + B\,C\,D) \cdot y \\
&= \Big[x + BCDA^{-1}x\Big] - BCDA^{-1}x \\
&= x
\end{aligned}
$$

As for every vector x,

$$(A + B\,C\,D)\Big[A^{-1} \;-\; A^{-1}B\Big(C^{-1} + D\,A^{-1}B\Big)^{-1}D\,A^{-1}\Big]x = x,$$

and by definition

$$AA^{-1}x = x,$$

it follows that

$$A^{-1} - A^{-1}B\Big(C^{-1} + D\,A^{-1}B\Big)^{-1}D\,A^{-1} = (A + B\,C\,D)^{-1}$$

If $\mathbf{A}$ is an $n \times n$ diagonal matrix, then computing $\mathbf{A}^{-1}$ simply involves taking reciprocals of the diagonal entries. Even if n is very large, only n such operations are needed, eliminating the need for expensive factorizations or dense matrix inversions. As $\mathbf{A}^{-1}$ is known, its computational cost can be considered negligible.
Let $\mathbf{B}$ be an $n \times k$ matrix (with $n \gg k$) and $\mathbf{C}$ be an invertible $k \times k$ matrix. Instead of inverting the full $n \times n$ matrix $\mathbf{A} + \mathbf{BCD}$, the Woodbury identity can be applied which shifts the inversion problem to a smaller $k \times k$ matrix:

$$\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}.$$

As k is small compared to n—for example 3 or 10—this smaller system 3×3 or 10×10 is much cheaper to invert in comparison to the $n \times n$ system, especially if n runs into the thousands or millions.

In terms of computational complexity, a naive inversion of $\mathbf{A} + \mathbf{BCD}$ costs $O(n^3)$ when $n \gg k$. With the Woodbury identity and a known $\mathbf{A}^{-1}$, the cost instead involves:

- Multiplications with $\mathbf{B}$ and $\mathbf{D}$, which is on the order of $O(nk^2)$.
- The inversion of the $k \times k$ matrix $\mathbf{C}$, which costs $O(k^3)$.

Because k is smaller in comparison to n, this leads to a computational advantage and results in computing the right-hand side of r.h.s of the formula.

3.

$$\mathcal{N}(x \mid \mu, \sigma^2) \;=\; \frac{1}{\sqrt{2\pi\,\sigma^2}}\;\exp\Bigl(-\frac{(x-\mu)^2}{2\,\sigma^2}\Bigr)$$

$$\mathcal{N}\Bigl(x \mid \mu, \tfrac{1}{\lambda}\Bigr) \;=\; \sqrt{\frac{\lambda}{2\pi}}\;\exp\Bigl(-\frac{\lambda}{2}(x-\mu)^2\Bigr)$$

$$\begin{aligned}
p(\mathbf{X} \mid \mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}}\;\exp\Bigl(-\frac{\lambda}{2}(x_n-\mu)^2\Bigr) \\
&=\; \Bigl(\sqrt{\frac{\lambda}{2\pi}}\Bigr)^N\;\exp\Bigl(-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\Bigr) \\
&=\; \left(\frac{\lambda}{2\pi}\right)^{\frac{N}{2}}\;\exp\Bigl(-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\Bigr) \\
&=\; \lambda^{\frac{N}{2}}\,(2\pi)^{-\frac{N}{2}}\;\exp\Bigl(-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\Bigr).
\end{aligned}$$

$$p(\mathbf{X} \mid \mu, \lambda) \propto \lambda^{\frac{N}{2}}\;\exp\Bigl(-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\Bigr).$$

$$p(\mu, \lambda) = \mathcal{N}\bigl(\mu \mid \mu_0,\;(\beta\,\lambda)^{-1}\bigr)\;\mathrm{Gam}\bigl(\lambda \mid a,\;b\bigr)$$

$$\mathcal{N}\bigl(\mu \mid \mu_0,\;(\beta\,\lambda)^{-1}\bigr) = \sqrt{\frac{\beta\,\lambda}{2\pi}}\;\exp\Bigl(-\frac{\beta\,\lambda}{2}(\mu-\mu_0)^2\Bigr)$$

$$\mathrm{Gam}\bigl(\lambda \mid a,\;b\bigr) = \frac{b^a}{\Gamma(a)}\,\lambda^{a-1}\;\exp(-b\,\lambda)$$

$$\begin{aligned}
p(\mu, \lambda) &= \sqrt{\frac{\beta\,\lambda}{2\pi}}\;\exp\Bigl(-\frac{\beta\,\lambda}{2}(\mu-\mu_0)^2\Bigr)\cdot\frac{b^a}{\Gamma(a)}\,\lambda^{a-1}\;\exp(-b\,\lambda) \\
&= \sqrt{\frac{\beta}{2\pi}}\,\lambda^{\frac{1}{2}}\cdot\frac{b^a}{\Gamma(a)}\,\lambda^{a-1}\cdot\exp\Bigl(-\frac{\beta\,\lambda}{2}(\mu-\mu_0)^2 - b\,\lambda\Bigr) \\
&= \sqrt{\frac{\beta}{2\pi}}\,\frac{b^a}{\Gamma(a)}\,\lambda^{\frac{1}{2}+(a-1)}\exp\Bigl(-\frac{\beta\,\lambda}{2}(\mu-\mu_0)^2 - b\,\lambda\Bigr) \\
&= \sqrt{\frac{\beta}{2\pi}}\,\frac{b^a}{\Gamma(a)}\,\lambda^{a-\frac{1}{2}}\exp\Bigl(-\frac{\beta\,\lambda}{2}(\mu-\mu_0)^2 - b\,\lambda\Bigr) \\
&= \sqrt{\frac{\beta}{2\pi}}\,\frac{b^a}{\Gamma(a)}\,\lambda^{a-\frac{1}{2}}\exp\Bigl(-\frac{\beta\,\lambda}{2}\,(\mu^2 - 2\mu\,\mu_0 + \mu_0^2) - b\,\lambda\Bigr) \\
&= \sqrt{\frac{\beta}{2\pi}}\,\frac{b^a}{\Gamma(a)}\,\lambda^{a-\frac{1}{2}}\exp\Bigl(-\frac{\beta\,\lambda}{2}\,\mu^2 + \beta\,\lambda\,\mu\,\mu_0 - \frac{\beta\,\lambda}{2}\,\mu_0^2 - b\,\lambda\Bigr)
\end{aligned}$$

$$p(\mu,\lambda) = \sqrt{\frac{\beta}{2\pi}}\,\frac{b^a}{\Gamma(a)}\,\lambda^{a-\frac{1}{2}}\exp\Bigl(-\frac{\beta\,\lambda\mu^2}{2}\;+\beta\,\mu_0\lambda\,\mu\;-\;\frac{\beta\,\lambda}{2}\,\mu_0^2 - b\,\lambda\Bigr) \quad\longrightarrow (1)$$

Equation (2.153):

$$\begin{aligned}
p(\mu,\lambda) &\propto \Bigl[\lambda^{\frac{1}{2}}\exp\Bigl(-\tfrac{\lambda\mu^2}{2}\Bigr)\Bigr]^{\beta}\;\exp\Bigl(c\,\lambda\,\mu\;-\;d\,\lambda\Bigr)\\
&= \Bigl[\lambda^{\frac{1}{2}}\Bigr]^{\beta}\;\Bigl[\exp\Bigl(-\tfrac{\lambda\mu^2}{2}\Bigr)\Bigr]^{\beta}\;\exp\Bigl(c\,\lambda\,\mu\;-\;d\,\lambda\Bigr)\\
&= \lambda^{\frac{\beta}{2}}\exp\Bigl(-\tfrac{\beta\,\lambda\,\mu^2}{2}\Bigr)\exp\Bigl(c\,\lambda\,\mu\;-\;d\,\lambda\Bigr)\\
&= \lambda^{\frac{\beta}{2}}\exp\Bigl(-\tfrac{\beta\,\lambda\,\mu^2}{2}\;+\;c\,\lambda\,\mu\;-\;d\,\lambda\Bigr)\quad\longrightarrow (2).
\end{aligned}$$

From (1) and (2):

From the power of λ : $\quad a - \dfrac{1}{2} = \dfrac{\beta}{2} \quad\Longrightarrow\quad a = \dfrac{\beta}{2} + \dfrac{1}{2} = \dfrac{\beta+1}{2}.$

From the $\lambda\mu$ term: $\quad \beta\,\mu_0 = c \quad\Longrightarrow\quad \mu_0 = \dfrac{c}{\beta}.$

From the constant term in the exponent: $\quad -\dfrac{\beta\,\mu_0^2}{2} - b = -d \quad\Longrightarrow\quad d = b + \dfrac{\beta\,\mu_0^2}{2}.$

Substituting $\mu_0 = \frac{c}{\beta}$ gives $\frac{\beta\,\mu_0^2}{2} = \frac{\beta}{2}\left(\frac{c}{\beta}\right)^2 = \frac{c^2}{2\beta}$, so that$b = d - \frac{c^2}{2\beta}$.

$$\lambda^{\beta/2} = \lambda^{a-\frac{1}{2}} \quad\Longrightarrow\quad a - \frac{1}{2} = \frac{\beta}{2} \quad\Longrightarrow\quad a = \frac{1+\beta}{2}$$

The parameters of the distribution are:

$$\mu_0 = \frac{c}{\beta}, \quad a = \frac{\beta+1}{2}, \quad b = d - \frac{c^2}{2\beta}$$

$$\begin{aligned}
&p(\mu,\lambda\mid\mathbf{X}) \;\;\propto\;\; p(\mathbf{X}\mid\mu,\lambda)\;\cdot\;p(\mu,\lambda)\\
&\propto\;\lambda^{\frac{N}{2}}\exp\Bigl(-\tfrac{\lambda}{2}\sum_{n=1}^{N}\bigl(x_n-\mu\bigr)^2\Bigr)\cdot\lambda^{\frac{\beta}{2}}\;\exp\Bigl(-\tfrac{\beta\,\lambda\,\mu^2}{2}\;+\;c\,\lambda\,\mu\;-\;d\,\lambda\Bigr)\\
&\propto\;\lambda^{\frac{N+\beta}{2}}\exp\bigl[-\tfrac{\lambda}{2}\bigl(\sum_{n=1}^{N}\bigl(x_n-\mu\bigr)^2+\beta\mu^2\bigr)+c\lambda\mu-d\lambda\bigr]\\
&\propto\;\lambda^{\frac{N+\beta}{2}}\exp\bigl[-\tfrac{\lambda}{2}\bigl(\sum_{n=1}^{N}x_n^2\;-\;2\,\mu\sum_{n=1}^{N}x_n\;+\;N\,\mu^2+\beta\mu^2\bigr)+c\lambda\mu-d\lambda\bigr]\\
&\propto\;\lambda^{\frac{N+\beta}{2}}\exp\Bigl[-\frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\;-\;d\,\lambda\;+\;\lambda\,\mu\,\bigl(\sum_{n=1}^{N}x_n+c\bigr)\;-\;\frac{\lambda}{2}(N+\beta)\,\mu^2\Bigr]\\
&\propto\;\lambda^{\frac{N+\beta}{2}}\exp\Bigl[-\tfrac{\lambda(N+\beta)\mu^2}{2}\;+\;\Bigl(c+\sum_{n=1}^{N}x_n\Bigr)\lambda\,\mu\;-\;\Bigl(d+\tfrac{1}{2}\sum_{n=1}^{N}x_n^2\Bigr)\lambda\Bigr]
\end{aligned}$$

$$d' = d + \tfrac{1}{2}\sum_{n=1}^{N}x_n^2$$

$$\mu_0' = \tfrac{c'}{\beta'} = \frac{c + \sum_{n=1}^{N} x_n}{N+\beta}$$

$$a' \;=\; \frac{\beta + N + 1}{2}$$

$$b' = d' - \frac{c'^2}{2\,\beta'}$$

By matching exponents and powers of λ and μ, the posterior distribution is the same functional form as the prior distribution but with updated parameters. Under the Normal–Gamma parameterization:

$$p(\mu, \lambda \mid \mathbf{X}) \;=\; \mathcal{N}\big(\mu \mid \mu_0', (\beta' \lambda)^{-1}\big)\; \Gamma(\lambda \mid a', b'),$$

where

$$\beta' = \beta + N, \quad c' = c + \sum_{n=1}^{N} x_n, \mu_0' = \tfrac{c'}{\beta'} = \frac{c + \sum_{n=1}^{N} x_n}{N+\beta}, \quad a' \;=\; \frac{\beta + N + 1}{2} \quad b' \;=\; d' - \frac{c'^2}{2\,\beta'}.$$

The posterior distribution is also a Gaussian–Gamma distribution of the same functional form as the prior, but with updated parameters, confirming that the posterior remains a Gaussian–Gamma distribution of the same functional form as the prior.

4.

a.

Wishart as a conjugate prior to $\Lambda = \Sigma^{-1}$ for Gaussian distribution $\mathcal{N}(\mu, \Lambda^{-1})$

A Wishart distribution over the precision matrix $\Lambda \in \mathbb{R}^{D\times D}$ with parameters (W, ν) is given by:

$$\mathcal{W}(\Lambda \mid W, \nu) \;=\; B(W, \nu)\, |\Lambda|^{\frac{\nu - D - 1}{2}} \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(W^{-1}\Lambda)\Big] \longrightarrow (1)$$

$$\mathcal{W}(\Lambda \mid W, \nu) \;\propto\; |\Lambda|^{\frac{\nu - D - 1}{2}} \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(W^{-1}\Lambda)\Big]$$

$$p(X \mid \mu, \Lambda) \;\propto\; |\Lambda|^{\frac{N}{2}} \exp\!\Big[-\tfrac{1}{2}\sum_{n=1}^{N} (x_n - \mu)^T \Lambda (x_n - \mu)\Big]$$

Let $S = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)\,(x_n - \mu)^T$.Then $\sum_{n=1}^{N} (x_n - \mu)^T \Lambda (x_n - \mu) = N\,\mathrm{Tr}(\Lambda\, S)$.

$$p(X \mid \mu, \Lambda) \;\propto\; |\Lambda|^{\frac{N}{2}} \exp\!\big[-\tfrac{1}{2}\,\mathrm{Tr}(N\,S\,\Lambda)\big] \;=\; |\Lambda|^{\frac{N}{2}} \exp\!\big[-\tfrac{N}{2}\,\mathrm{Tr}(\Lambda S)\big]$$

Posterior in Λ:

$$\begin{aligned}
p(\Lambda \mid X, W, \nu) \;&\propto\; p(X \mid \mu, \Lambda)\, \mathcal{W}(\Lambda \mid W, \nu) \\
&= |\Lambda|^{\frac{N}{2}} \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(N\,S\,\Lambda)\Big] \times |\Lambda|^{\frac{\nu - D - 1}{2}} \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(W^{-1}\Lambda)\Big] \\
&= |\Lambda|^{\frac{N}{2}} \;\times\; |\Lambda|^{\frac{\nu - D - 1}{2}} \cdot \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(N\,S\,\Lambda)\Big] \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(W^{-1}\Lambda)\Big] \\
&= |\Lambda|^{\frac{N + \nu - D - 1}{2}} \cdot \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}\big((W^{-1} + N\,S)\,\Lambda\big)\Big] \longrightarrow (2)
\end{aligned}$$

From (1) and (2),

$$\mathcal{W}\big(\Lambda \mid (W^{-1} + N\,S)^{-1},\, N + \nu\big) = |\Lambda|^{\frac{N + \nu - D - 1}{2}} \exp\!\Big(-\tfrac{1}{2}\,\mathrm{Tr}\big((W^{-1} + N\,S)\,\Lambda\big)\Big)$$

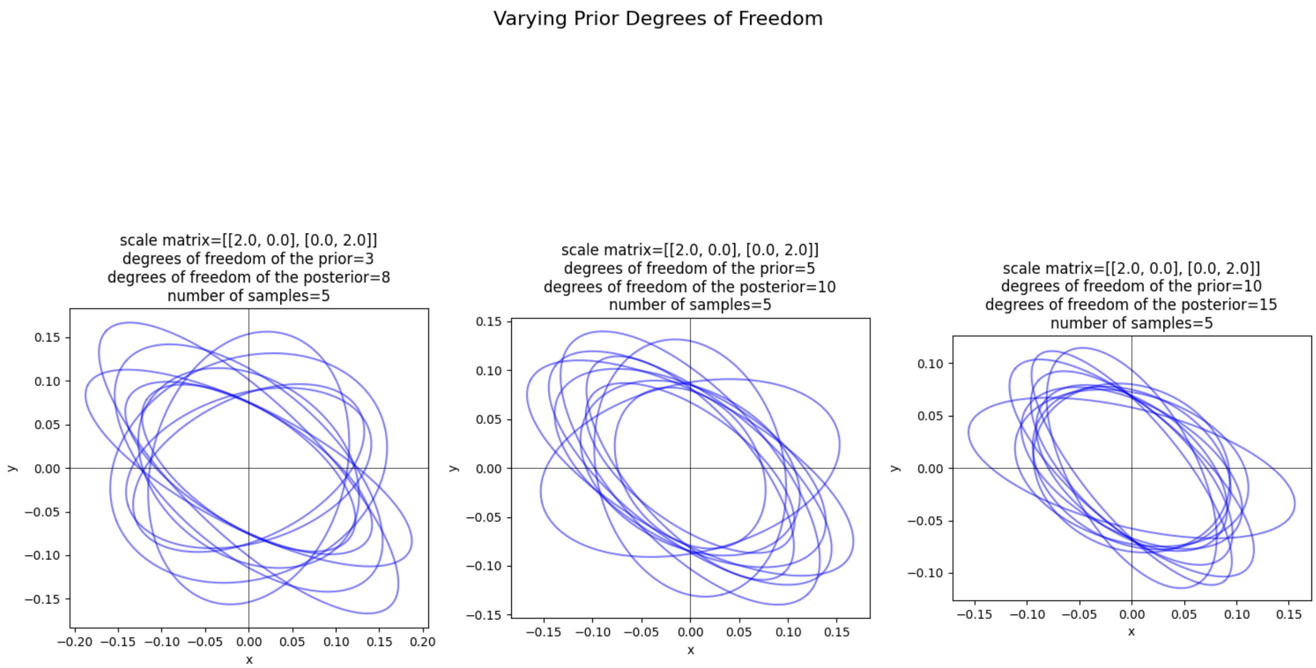
$$\mathcal{W}(\Lambda \mid W, \nu) \;=\; B(W, \nu)\, |\Lambda|^{\frac{\nu - D - 1}{2}} \exp\!\Big[-\tfrac{1}{2}\,\mathrm{Tr}(W^{-1}\Lambda)\Big]$$

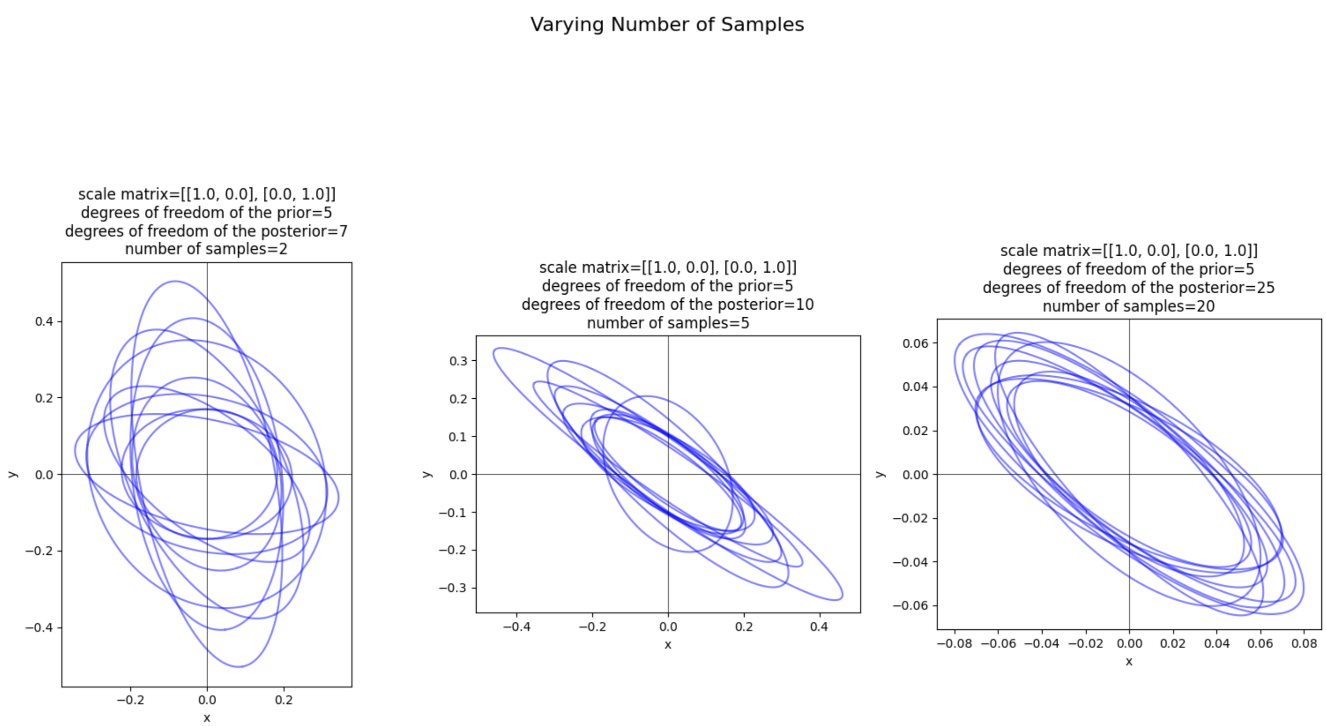
Therefore, the posterior is in the same form as a Wishart distribution with parameters:

$$\nu = N + \nu, \quad W = (W^{-1} + NS)^{-1}$$

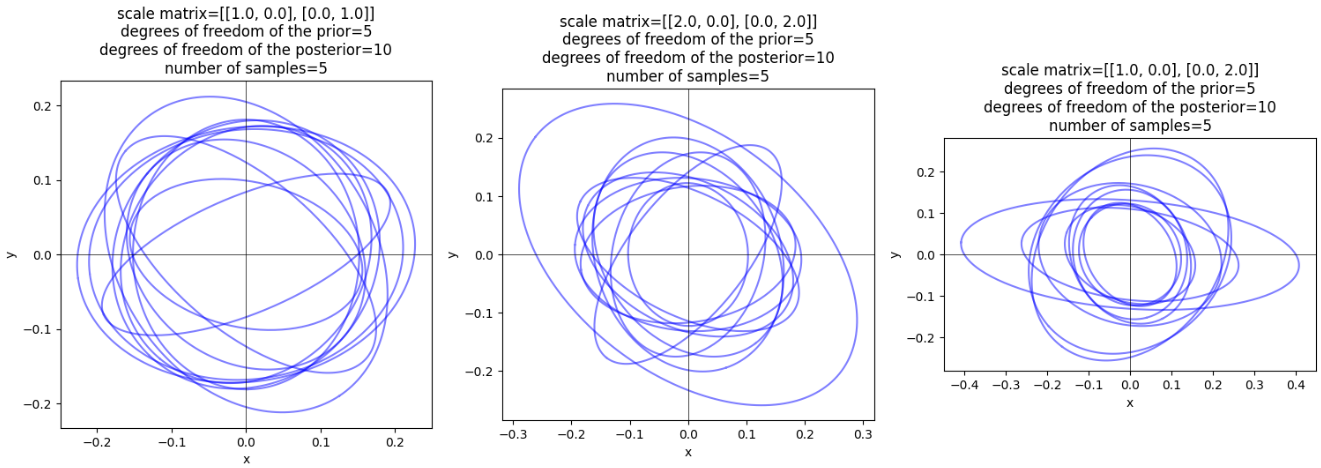
As $\mathcal{W}(\mathbf{\Lambda} \mid W, \nu)$ (Wishart Prior) $\times\; p(X \mid \boldsymbol{\mu}, \mathbf{\Lambda})$ (Gaussian likelihood) $= \mathcal{W}(\mathbf{\Lambda} \mid W, \nu)$ (Wishart posterior), this closure property under posterior updates defines a conjugate prior for the Gaussian precision matrix.

b.


Varying Prior Degrees of Freedom
scale matrix=[[2.0, 0.0], [0.0, 2.0]]
degrees of freedom of the prior=3
degrees of freedom of the posterior=8
number of samples=5
scale matrix=[[2.0, 0.0], [0.0, 2.0]]
degrees of freedom of the prior=5
degrees of freedom of the posterior=10
number of samples=5
scale matrix=[[2.0, 0.0], [0.0, 2.0]]
degrees of freedom of the prior=10
degrees of freedom of the posterior=15
number of samples=5
x
y


Varying Number of Samples
scale matrix=[[1.0, 0.0], [0.0, 1.0]]
degrees of freedom of the prior=5
degrees of freedom of the posterior=7
number of samples=2
scale matrix=[[1.0, 0.0], [0.0, 1.0]]
degrees of freedom of the prior=5
degrees of freedom of the posterior=10
number of samples=5
scale matrix=[[1.0, 0.0], [0.0, 1.0]]
degrees of freedom of the prior=5
degrees of freedom of the posterior=25
number of samples=20
x
y

Varying Scale Matrix




```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import wishart

def generate_data(n, cov):
    return np.random.multivariate_normal(mean=[0, 0], cov=cov, size=n)

def sample_wishart(df, scale, num_samples=5):
    samples = []
    for _ in range(num_samples):
        W = wishart.rvs(df=df, scale=scale)
        samples.append(W)
    return samples

def plot_precision_ellipses(precisions, ax, title):
    for p in precisions:
        eigvals, eigvecs = np.linalg.eigh(p)
        if np.any(eigvals <= 0):
            continue

        angles = np.linspace(0, 2*np.pi, 200)
        circle = np.stack([np.cos(angles), np.sin(angles)], axis=1)
        scale_matrix = np.diag(1.0 / np.sqrt(eigvals))
        ellipse_y = circle @ scale_matrix
        ellipse_x = ellipse_y @ eigvecs.T
        ax.plot(ellipse_x[:, 0], ellipse_x[:, 1], 'b', alpha=0.5)

    ax.set_aspect('equal', 'box')
    ax.set_title(title)
    ax.set_xlabel('x')
    ax.set_ylabel('y')
    ax.axhline(0, color='black', linewidth=0.5)
    ax.axvline(0, color='black', linewidth=0.5)

def variation_1_distribution():
    n = 5
    scale_matrix = np.eye(2) * 2.0
```

```python
    cov = np.array([[1.0, 0.5], [0.5, 1.0]])
    X = generate_data(n, cov)
    sum_T = X.T @ X
    dfs = [3, 5, 10]

    fig, axes = plt.subplots(1, 3, figsize=(15, 4))

    for i, df_prior in enumerate(dfs):
        df_post = df_prior + n
        scale_post = scale_matrix + sum_T
        posterior_samples = sample_wishart(df_post, scale_post, num_samples=10)

        title = (
            f"scale matrix={scale_matrix.tolist()}\n"
            f"degrees of freedom of the prior={df_prior}\n"
            f"degrees of freedom of the posterior={df_post}\n"
            f"number of samples={n}"
        )
        plot_precision_ellipses(posterior_samples, axes[i], title)

    fig.suptitle("Varying Prior Degrees of Freedom", fontsize=16)
    plt.tight_layout()
    plt.show()

def variation_2_distribution():
    df_prior = 5
    scale_matrix = np.eye(2)
    cov = np.array([[1.0, 0.5],[0.5, 1.0]])
    ns = [2, 5, 20]
    fig, axes = plt.subplots(1, 3, figsize=(15, 4))

    for i, n in enumerate(ns):
        X = generate_data(n, cov)
        sum_T = X.T @ X
        df_post = df_prior + n
        scale_post = scale_matrix + sum_T
        posterior_samples = sample_wishart(df_post, scale_post, num_samples=10)

        title = (
            f"scale matrix={scale_matrix.tolist()}\n"
            f"degrees of freedom of the prior={df_prior}\n"
            f"degrees of freedom of the posterior={df_post}\n"
            f"number of samples={n}"
        )
        plot_precision_ellipses(posterior_samples, axes[i], title)

    fig.suptitle("Varying Number of Samples", fontsize=16)
    plt.tight_layout()
    plt.show()

def variation_3_distribution():
    df_prior = 5
    n = 5
    scale_matrix_list = [np.eye(2), 2.0 * np.eye(2), np.diag([1.0, 2.0])]
    cov = np.array([[1.0, 0.5],[0.5, 1.0]])
    X = generate_data(n, cov)
    sum_T = X.T @ X

    fig, axes = plt.subplots(1, 3, figsize=(15, 4))

    for i, scale_matrix in enumerate(scale_matrix_list):
        df_post = df_prior + n
        scale_post = scale_matrix + sum_T
        posterior_samples = sample_wishart(df_post, scale_post, num_samples=10)

        title = (
            f"scale matrix={scale_matrix.tolist()}\n"
            f"degrees of freedom of the prior={df_prior}\n"
            f"degrees of freedom of the posterior={df_post}\n"
            f"number of samples={n}"
        )
```

```python
        plot_precision_ellipses(posterior_samples, axes[i], title)

    fig.suptitle("Varying Scale Matrix", fontsize=16)
    plt.tight_layout()
    plt.show()

def main():
    variation_1_distribution()
    variation_2_distribution()
    variation_3_distribution()

if __name__ == "__main__":
    main()
```

5.

$$
\begin{aligned}
\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{D}{2}}\, |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\,(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\Big\} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\Big(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;-\; 2\,\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;+\; \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big)\Big\} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\,\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;+\; \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;-\; \tfrac{1}{2}\,\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big\} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\,\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;+\; \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\Big\}\; \exp\Big\{-\tfrac{1}{2}\,\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big\} \\
&= (2\pi)^{-\frac{D}{2}} \cdot |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\tfrac{1}{2}\,\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big\}\; \exp\Big\{-\tfrac{1}{2}\,\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} \;+\; \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\Big\}
\end{aligned}
$$

Exponential-family form:

$$
p(\mathbf{x} \mid \boldsymbol{\eta}) \;=\; h(\mathbf{x}) \;\cdot\; g(\boldsymbol{\eta}) \;\cdot\; \exp\!\big(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\big)
$$

$$
\begin{aligned}
h(\mathbf{x}) \;&=\; (2\pi)^{-\frac{D}{2}} \\
\mathbf{u}(\mathbf{x}) \;&=\; \big(\mathbf{x},\, \mathbf{x}\mathbf{x}^\top\big) \quad \text{(which generalizes the univariate } (x, x^2)\text{)} \\
\boldsymbol{\eta} \;&=\; \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} \;=\; \begin{pmatrix} \Sigma^{-1}\,\boldsymbol{\mu} \\ -\tfrac{1}{2}\,\Sigma^{-1} \end{pmatrix}
\end{aligned}
$$

Solving for $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ in terms of $\boldsymbol{\eta}$.

$$
\boldsymbol{\eta}_2 = -\frac{1}{2}\,\Sigma^{-1} \quad\Longrightarrow\quad \Sigma^{-1} = -2\,\boldsymbol{\eta}_2 \quad\Longrightarrow\quad \Sigma = \big(\Sigma^{-1}\big)^{-1} = \big(-2\,\boldsymbol{\eta}_2\big)^{-1} = -\tfrac{1}{2}\,\boldsymbol{\eta}_2^{-1}.
$$

$$
\boldsymbol{\eta}_1 = \Sigma^{-1}\,\boldsymbol{\mu} \quad\Longrightarrow\quad \boldsymbol{\mu} = \Sigma\,\boldsymbol{\eta}_1 = \Big(-\tfrac{1}{2}\,\boldsymbol{\eta}_2^{-1}\Big)\,\boldsymbol{\eta}_1
$$

$$
|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \;=\; \big|-2\,\boldsymbol{\eta}_2\big|^{\frac{D}{2}} \quad \text{(for } D \text{ dimensions)}
$$

$$
g(\boldsymbol{\eta}) \;=\; |\Sigma|^{-\frac{1}{2}}\; \exp\!\Big(-\tfrac{1}{2}\,\boldsymbol{\mu}^\top \Sigma^{-1}\,\boldsymbol{\mu}\Big)
$$

$$\boldsymbol{\mu}^\top \Sigma^{-1} \boldsymbol{\mu} \;=\; (\Sigma\, \boldsymbol{\eta}_1)^\top \Sigma^{-1} (\Sigma\, \boldsymbol{\eta}_1) \;=\; \boldsymbol{\eta}_1^\top \Sigma\, \boldsymbol{\eta}_1 \;=\; \boldsymbol{\eta}_1^\top \Big(-\tfrac{1}{2}\, \boldsymbol{\eta}_2^{-1}\Big) \boldsymbol{\eta}_1 \;=\; -\tfrac{1}{2}\, \boldsymbol{\eta}_1^\top\, \boldsymbol{\eta}_2^{-1}\, \boldsymbol{\eta}_1$$

Therefore,

$$\begin{aligned} g(\boldsymbol{\eta}) \;&=\; \big|\Sigma\big|^{-\frac{1}{2}}\; \exp\Big(-\tfrac{1}{2}\, \boldsymbol{\mu}^\top\, \Sigma^{-1}\, \boldsymbol{\mu}\Big) \\ &= \big|\Sigma\big|^{-\frac{1}{2}}\; \exp\Big(-\tfrac{1}{2}\, \boldsymbol{\eta}_1^\top\, \boldsymbol{\eta}_2^{-1}\, \boldsymbol{\eta}_1\Big) \\ &= \big|-2\, \boldsymbol{\eta}_2\big|^{\frac{D}{2}}\; \exp\Big(\tfrac{1}{4}\, \boldsymbol{\eta}_1^\top\, \boldsymbol{\eta}_2^{-1}\, \boldsymbol{\eta}_1\Big) \end{aligned}$$

Univariate ($D = 1$) case:

$$g(\boldsymbol{\eta}) \;=\; \big|-2\, \eta_2\big|^{\frac{1}{2}}\; \exp\Big(\tfrac{\eta_1^2}{4\, \eta_2}\Big)$$

Multivariate ($D > 1$) case:

$$g(\boldsymbol{\eta}) \;=\; \big|-2\, \boldsymbol{\eta}_2\big|^{\frac{D}{2}}\; \exp\Big(\tfrac{1}{4}\, \boldsymbol{\eta}_1^\top\, \boldsymbol{\eta}_2^{-1}\, \boldsymbol{\eta}_1\Big)$$