Multivariate Gaussian Distribution, Bayesian Linear Regression, Feature Map, Kernel Trick

Kaiwen Zhou

Contents

1 Useful Results in Matrix Algebra

Proposition 1.1. ***(Matrix Identities)*** *For square invertible matrices Σ and $\mathbf{S}$ such that $\mathbf{S}+\Sigma$ is invertible, we have*

(a) $\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}$.

(b) $\mathbf{S}^{-1} - \mathbf{S}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{S}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}$.

(c) $\mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{S}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}$

(d) $\mathbf{S}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}$

Solution: For (a) and (b), by symmetry, we just have to prove (a), and we have

$$\begin{aligned}
&\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1} \Longleftarrow \mathbf{I} - \mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{\Sigma} \\
&\Longleftarrow \mathbf{S}^{-1}+\mathbf{\Sigma}^{-1}-\mathbf{\Sigma}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{\Sigma}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1}) \Longleftarrow (\mathbf{S}+\mathbf{\Sigma})\mathbf{S}^{-1} = \mathbf{\Sigma}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1}) \\
&\Longleftarrow \mathbf{I}+\mathbf{\Sigma}\mathbf{S}^{-1} = \mathbf{\Sigma}\mathbf{S}^{-1}+\mathbf{I} \Longleftarrow 0 = 0
\end{aligned}$$

For (c) and (d), by symmetry, we just have to prove (c), and we have

$$\begin{aligned}
&\mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{S}^{-1} = (\mathbf{S}+\mathbf{\Sigma})^{-1} \Longleftarrow (\mathbf{S}+\mathbf{\Sigma})\mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1} = \mathbf{S} \Longleftarrow (\mathbf{S}+\mathbf{\Sigma})\mathbf{\Sigma}^{-1} = \mathbf{S}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1}) \\
&\Longleftarrow \mathbf{S}\mathbf{\Sigma}^{-1}+\mathbf{I} = \mathbf{I}+\mathbf{S}\mathbf{\Sigma}^{-1} \Longleftarrow 0 = 0
\end{aligned}$$

Therefore, the proposition is proved.

Proposition 1.2. ***(Block Inversion)*** *If a matrix is partitioned into four blocks, it can be inverted blockwise as follows:*

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}$$

where $\mathbf{A}$ and $\mathbf{D}$ are square blocks of arbitrary size, and $\mathbf{B}$ and $\mathbf{C}$ are conformable with them for partitioning. Furthermore, $\mathbf{D}$ and the Schur complement of $\mathbf{D}$ in $\mathbf{P}$: $\mathbf{P}/\mathbf{D} = \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ must be invertible.

2 Topic: Multivariate Gaussian Distributions

Problem 2.1. *Suppose $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}^\top$ is multivariate Gaussian distributed, i.e.*

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^\top & \mathbf{\Sigma}_{22} \end{bmatrix} \right)$$

Prove the following properties.

(a) *The marginal of $\mathbf{x}_1$ is Gaussian, that is $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$.*

Solution: By definition, every entry of $\mathbf{x}_1$ is Gaussian. Therefore, every linear combination of entries of $\mathbf{x}_1$ is also Gaussian, and hence, by equivalent definition of multivariate normal distribution, $\mathbf{x}_1$ is multivariate normally distributed.

Additionally, the multivariate normal distribution is determined exclusively by its mean and covariance matrix - which can be directly found in the provided matrix, namely, μ_1 and Σ_{11}.

Therefore, $\mathbf{x}_1$ is multivariate Gaussian distributed, and

$$\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$$

Alternative Solution: Suppose $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x}_1 \in \mathbb{R}^d$. Set $\mathbf{A} = \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{d\times n}$ and $\mathbf{b} = \mathbf{0} \in \mathbb{R}^d$ in Item (e), we have

$$\mathbb{E}[\mathbf{x}_1] = \mathbb{E}[\mathbf{A}\mathbf{x}+\mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{x}]+\mathbf{b} = \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \boldsymbol{\mu}_1$$

and

$$Cov[\mathbf{x}_1, \mathbf{x}_1] = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top = \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^\top & \mathbf{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \end{bmatrix}^\top = \mathbf{\Sigma}_{11}$$

Therefore, $\mathbf{x}_1$ is multivariate Gaussian distributed, and

$$\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$$

(b) *$\mathbf{x}_1$ conditional on $\mathbf{x}_2$ is Gaussian*

$$\mathbf{x}_1 \mid \mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \mathbf{\Sigma}_{11|2})$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{1|2} &:= \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2) \\
\mathbf{\Sigma}_{11|2} &:= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^\top
\end{aligned}$$

Solution: Suppose $\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^\top & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix}A & B\\ C & D\end{bmatrix}$ where A, D are symmetric and $C = B^\top$, we obtain

$$P = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^\top & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix}A & B\\ C & D\end{bmatrix} = \begin{bmatrix}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1} & -\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma_{12}^\top\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1} & \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{12}^\top\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{bmatrix}$$

Here, Σ_{22} and the Schur complement of Σ_{22} in P : $P/\Sigma_{22} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top$ must be invertible.

By definition, we have

$$\begin{aligned}
f_{x_1|x_2} &\propto \frac{f_{x_1,x_2}}{f_{x_2}}\\
&\propto \frac{\exp\left(-\frac{1}{2}\left[\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^\top\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^\top & \Sigma_{22}\end{bmatrix}^{-1}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)\right]\right)}{\exp\left(-\frac{1}{2}\left[(x_2-\mu_2)^\top\Sigma_{22}^{-1}(x_2-\mu_2)\right]\right)}\\
&\propto \exp\left(-\frac{1}{2}\left[x_1^\top Ax_1 + x_1^\top(B(x_2-\mu_2)-A\mu_1)+\left((x_2-\mu_2)^\top C-\mu_1^\top A\right)x_1\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left[\left(x_1-A^{-1}(A\mu_1-B(x_2-\mu_2))\right)^\top A\left(x_1-A^{-1}(A\mu_1-B(x_2-\mu_2))\right)\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left[\left(x_1-(\mu_1-A^{-1}B(x_2-\mu_2))\right)^\top A\left(x_1-(\mu_1-A^{-1}B(x_2-\mu_2))\right)\right]\right)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mu_{1|2} &= \mu_1 - A^{-1}B(x_2-\mu_2) = \mu_1 - \left[\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1}\right]^{-1}\left[-\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\right] = \mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\\
\Sigma_{11|2} &= A^{-1} = \left[\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top\right)^{-1}\right]^{-1} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^\top
\end{aligned}$$

(c) *How do you interpret (1)-(2) in words?*

Solution:

Given, in this case, x_2 is known, we will use this information to further our understanding on x_1.

Judging by how much the observation x_2 is distant from its mean μ_2, we modify the mean μ_1 of x_1 through the covariance matrix Σ_{12}.

By observing the values of x_2, we gain new knowledge or understanding. This should lead to a "better" understanding of x_1, resulting in a narrower range of potential values for x_1. Consequently, the corresponding covariance matrix will decrease by the specified amount.

(d) *Suppose that $x \sim \mathcal{N}(\mu, \Sigma)$ and $y \sim \mathcal{N}(m, S)$ are independently Gaussian distributed random vectors (of the same dimension). Show that their sum $z = x + y$ is Gaussian with a PDF given by the convolution of the individual densities, i.e. $z \sim \mathcal{N}(\mu + m, \Sigma + S)$.*

Solution: The convolution of the individual densities gives

$$\begin{aligned}
f_z(z|m,\mu,S,\Sigma) &= \int f_y(z-w|m,S)\cdot f_x(w|\mu,\Sigma)dw\\
&= \int \frac{1}{\sqrt{(2\pi)^n|S|}}\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\left[(w-m)^\top S^{-1}(w-m)+(z-w-\mu)^\top\Sigma^{-1}(z-w-\mu)\right]\right)dw\\
&= \int \frac{1}{\sqrt{(2\pi)^n|S|}}\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\\
&\quad\cdot\exp\left(-\frac{1}{2}\left\{\left[w^\top-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})(S^{-1}+\Sigma^{-1})^{-1}\right](S^{-1}+\Sigma^{-1})\left[w^\top-(S^{-1}+\Sigma^{-1})^{-1}(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})\right]\right\}\right.\\
&\quad\left.-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})(S^{-1}+\Sigma^{-1})(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})+m^\top S^{-1}m+z^\top\Sigma^{-1}z+\mu^\top\Sigma^{-1}\mu-\mu^\top\Sigma^{-1}z-z^\top\Sigma^{-1}\mu\right)dw\\
&= \frac{\sqrt{(2\pi)^n|(S^{-1}+\Sigma^{-1})^{-1}|}}{\sqrt{(2\pi)^n|S|}\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\left\{-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})(S^{-1}+\Sigma^{-1})(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})\right.\right.\\
&\quad\left.\left.+m^\top S^{-1}m+z^\top\Sigma^{-1}z+\mu^\top\Sigma^{-1}\mu-\mu^\top\Sigma^{-1}z-z^\top\Sigma^{-1}\mu\right\}\right)\\
&\quad\cdot\int\frac{1}{\sqrt{(2\pi)^n|S^{-1}+\Sigma^{-1}|}}\exp\left(-\frac{1}{2}\left\{\left[w^\top-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})\right](S^{-1}+\Sigma^{-1})\left[w^\top-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})\right]\right\}\right)dw\\
&= \frac{\sqrt{(2\pi)^n|(S^{-1}+\Sigma^{-1})^{-1}|}}{\sqrt{(2\pi)^n|S|}\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\left\{-(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})(S^{-1}+\Sigma^{-1})(m^\top S^{-1}+z^\top\Sigma^{-1}-\mu^\top\Sigma^{-1})\right.\right.\\
&\quad\left.\left.+m^\top S^{-1}m+z^\top\Sigma^{-1}z+\mu^\top\Sigma^{-1}\mu-\mu^\top\Sigma^{-1}z-z^\top\Sigma^{-1}\mu\right\}\right)\cdot 1\\
&= \frac{\sqrt{(2\pi)^n|(S^{-1}+\Sigma^{-1})^{-1}|}}{\sqrt{(2\pi)^n|S|}\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\left\{z^\top(\Sigma^{-1}-\Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1})z+z^\top\lambda+\lambda^\top z\right.\right.\\
&\quad\left.\left.+\mu^\top\Sigma^{-1}\mu+m^\top S^{-1}m-(S^{-1}m-\Sigma^{-1}\mu)^\top(S^{-1}+\Sigma^{-1})^{-1}(S^{-1}m-\Sigma^{-1}\mu)\right\}\right)
\end{aligned}$$

where $\lambda = -\Sigma^{-1}\mu - \Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}S^{-1}m + \Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mu$. Then, use Proposition 1.1, we have $\Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1} = \Sigma^{-1}-(\Sigma+S)^{-1}$ and $\Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}S^{-1} = (\Sigma+S)^{-1}\Sigma S^{-1}$. Therefore, we obtain

$$(S+\Sigma)\lambda = (S+\Sigma)((\Sigma+S)^{-1}\Sigma S^{-1}m - S^{-1}m - (\Sigma+S)^{-1}\mu) = -\mu - m$$

and

$$\begin{aligned}
&\mu^\top\Sigma^{-1}\mu+m^\top S^{-1}m-(S^{-1}m-\Sigma^{-1}\mu)^\top(S^{-1}+\Sigma^{-1})^{-1}(S^{-1}m-\Sigma^{-1}\mu)\\
&= \mu^\top(\Sigma^{-1}-\Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1})\mu+m^\top(S^{-1}-S^{-1}(S^{-1}+\Sigma^{-1})^{-1}S^{-1})m-m^\top S^{-1}(S^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mu-\mu^\top\Sigma^{-1}(S^{-1}+\Sigma^{-1})^{-1}S^{-1}m\\
&= \mu^\top(S+\Sigma)^{-1}\mu+m^\top(S+\Sigma)^{-1}m-m^\top(S+\Sigma)^{-1}\mu-\mu^\top(S+\Sigma)^{-1}m\\
&= (m-\mu)^\top(S+\Sigma)^{-1}(m-\mu) = \lambda^\top(S+\Sigma)(S+\Sigma)^{-1}(S+\Sigma)\lambda
\end{aligned}$$

It follows that

$$\begin{aligned}
f_z &= \frac{\sqrt{(2\pi)^n|(S^{-1}+\Sigma^{-1})^{-1}|}}{\sqrt{(2\pi)^n|S|}\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\left\{(z-(\mu+m))^\top(S+\Sigma)^{-1}(z-(\mu+m))\right\}\right)\\
&= \frac{1}{\sqrt{(2\pi)^n}}\frac{\sqrt{|(S^{-1}+\Sigma^{-1})^{-1}|}}{\sqrt{|S|}\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\left\{(z-(\mu+m))^\top(S+\Sigma)^{-1}(z-(\mu+m))\right\}\right)
\end{aligned}$$

Since we have

$$\mathbf{S}\mathbf{\Sigma}^{-1} + \mathbf{I} = \mathbf{S}\mathbf{\Sigma}^{-1} + \mathbf{I} \Longrightarrow (\mathbf{S}+\mathbf{\Sigma})\mathbf{\Sigma}^{-1} = \mathbf{S}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1}) \Longrightarrow |(\mathbf{S}+\mathbf{\Sigma})||\mathbf{\Sigma}^{-1}| = |\mathbf{S}||(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})| \Longrightarrow \frac{|(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}|}{|\mathbf{S}||\mathbf{\Sigma}|} = \frac{1}{|\mathbf{S}+\mathbf{\Sigma}|}$$

Plug this in, we finally have

$$f_{\mathbf{z}}(\mathbf{z}|\mathbf{m},\boldsymbol{\mu},\mathbf{S},\mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{S}+\mathbf{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}\left\{(\mathbf{z}-(\boldsymbol{\mu}+\mathbf{m}))^\top(\mathbf{S}+\mathbf{\Sigma})^{-1}(\mathbf{z}-(\boldsymbol{\mu}+\mathbf{m}))\right\}\right) \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}+\mathbf{m},\mathbf{\Sigma}+\mathbf{S})$$

(e) Suppose $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$. *Define* $\mathbf{y} := \mathbf{A}\mathbf{x}+\mathbf{b}$ *where* $\mathbf{A} \in \mathbb{R}^{D\times d}$ *and* $\mathbf{b} \in \mathbb{R}^D$. *Show that* $\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top\right)$.

Solution: By linearity, we know $\mathbf{y}$ is multivariate Gaussian distribution. Now, we find its mean and covariance.

We have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x}+\mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{x}]+\mathbf{b} = \mathbf{A}\boldsymbol{\mu}+\mathbf{b}$$

and

$$Cov[\mathbf{y},\mathbf{y}] = \mathbb{E}[(\mathbf{A}\mathbf{x}+\mathbf{b}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})(\mathbf{A}\mathbf{x}+\mathbf{b}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\top] = \mathbf{A}\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\mathbf{A}^\top = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top$$

Hence, we have

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top\right).$$

3 Topic: Bayesian Estimation of the Mean (Known Covariance)

Problem 3.1. *Suppose* $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$, *where* $\mathbf{\Sigma}$ *is known. We want to infer the mean* $\boldsymbol{\mu}$ *from a set of observations* $\mathbf{X} := \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. *Assume the prior distribution is given by* $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0,\mathbf{\Sigma}_0)$. *Determine the posterior distribution* $p(\boldsymbol{\mu} \mid \mathbf{X})$.

Solution: By the Bayes' Formula, we have

$$p(\boldsymbol{\mu} \mid \mathbf{x}_1,\ldots,\mathbf{x}_N) \propto p(\mathbf{x}_1,\ldots,\mathbf{x}_N \mid \boldsymbol{\mu})p(\boldsymbol{\mu})$$

Since the likelihood function $p(\mathbf{x} \mid \boldsymbol{\mu})$ is Gaussian and is conjugate to the prior $p(\boldsymbol{\mu})$, which is also Gaussian, we get the posterior distribution is also Gaussian. Therefore, we only care about the mean and covariance of $\boldsymbol{\mu} \mid \mathbf{x}$. We obtain

$$\begin{aligned}
-2\log p(\boldsymbol{\mu} \mid \mathbf{x}_1,\ldots,\mathbf{x}_N) &\propto -2\log\left(p(\mathbf{x}_1,\ldots,\mathbf{x}_N \mid \boldsymbol{\mu})p(\boldsymbol{\mu})\right) \\
&\propto -2\log\left(\left[\prod_{i=1}^N p(\mathbf{x}_i \mid \boldsymbol{\mu})\right]p(\boldsymbol{\mu})\right) \\
&\propto (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \sum_{i=1}^N(\mathbf{x}_i-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top\mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) \\
&= \boldsymbol{\mu}^\top(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})\boldsymbol{\mu} - \boldsymbol{\mu}^\top\left(\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right)+\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) - \left(\left(\sum_{i=1}^N\mathbf{x}_i^\top\right)\mathbf{\Sigma}^{-1}+\boldsymbol{\mu}_0^\top\mathbf{\Sigma}_0^{-1}\right)\boldsymbol{\mu} + \boldsymbol{\mu}_0^\top\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{i=1}^N\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i \\
&\propto \left[\boldsymbol{\mu}-(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})^{-1}\left(\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right)+\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right]^\top(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})\left[\boldsymbol{\mu}-(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})^{-1}\left(\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right)+\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right]
\end{aligned}$$

Therefore, we obtain

$$\boldsymbol{\mu} \mid \mathbf{x}_1,\ldots,\mathbf{x}_N \sim \mathcal{N}\left((\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})^{-1}\left(\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right)+\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right),(\mathbf{\Sigma}_0^{-1}+N\mathbf{\Sigma}^{-1})^{-1}\right)$$

Alternative Form:

In Proposition 1.1 (c), substitute $\mathbf{S}$ with $\mathbf{\Sigma}_0^{-1}$ and $\mathbf{\Sigma}$ with $N\mathbf{\Sigma}^{-1}$; in Proposition 1.1 (d), substitute $\mathbf{S}$ with $N\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ with $\mathbf{\Sigma}_0^{-1}$, we obtain

$$(\mathbf{\Sigma}_0^{-1}+N^{-1}\mathbf{\Sigma}^{-1})^{-1} = \mathbf{\Sigma}_0(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}N^{-1}\mathbf{\Sigma} \quad \text{and} \quad (\mathbf{\Sigma}_0^{-1}+N^{-1}\mathbf{\Sigma}^{-1})^{-1} = N^{-1}\mathbf{\Sigma}(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}_0$$

$$\begin{aligned}
\Longrightarrow (\mathbf{\Sigma}_0^{-1}+N^{-1}\mathbf{\Sigma}^{-1})^{-1}\left(\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right)+\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) &= \mathbf{\Sigma}_0(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}N^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\left(\sum_{i=1}^N\mathbf{x}_i\right) + N^{-1}\mathbf{\Sigma}(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}_0\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \\
&= \mathbf{\Sigma}_0(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\left(\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i\right) + N^{-1}\mathbf{\Sigma}(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\boldsymbol{\mu}_0
\end{aligned}$$

Finally, we obtain

$$\boldsymbol{\mu} \mid \mathbf{x}_1,\ldots,\mathbf{x}_N \sim \mathcal{N}\left(\mathbf{\Sigma}_0(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\left(\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i\right)+N^{-1}\mathbf{\Sigma}(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\boldsymbol{\mu}_0, \quad N^{-1}\mathbf{\Sigma}(\mathbf{\Sigma}_0+N^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}_0\right)$$

4 Topic: Bayesian Estimation of the Mean (Unknown Variance)

Problem 4.1. *Consider a univariate Gaussian distribution* $\mathcal{N}\left(x \mid \mu,\tau^{-1}\right)$ *having the conjugate Gaussian gamma prior*

$$p(\mu,\lambda) := \mathcal{N}\left(\mu \mid \mu_0,(\beta\lambda)^{-1}\right)\mathrm{Gam}(\lambda \mid a,b),$$

where

$$\mathrm{Gam}(\tau \mid a,b) := \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau},$$

and $\Gamma(\alpha)$ *is the gamma function. Let* $\mathbf{x} = \{x_1,\ldots,x_N\}$ *denote a data set of i.i.d. observations. Prove that the posterior distribution is also a Gaussian-gamma distribution and determine the expressions for its parameters.*

Solution: The problem is saying that say we have a likelihood function $p(x|\mu,\lambda) \sim \mathcal{N}\left(\mu,\lambda^{-1}\right)$ and a prior function $p(\mu,\lambda) = p(\mu|\lambda)p(\lambda)$ where $\mu|\lambda \sim \mathcal{N}\left(\mu_0,(\beta\lambda)^{-1}\right)$ and $\lambda \sim \mathrm{Gam}(a,b)$, prove that the posterior $p(\mu,\lambda|x) \propto p(x|\mu,\lambda)p(\mu,\lambda)$ is also a Gaussian-Gamma distribution.

First, we note the Gaussian-Gamma prior is

$$p(\mu,\lambda) = p(\mu|\lambda)p(\lambda) = \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}}\exp\left(-\frac{(\mu-\mu_0)^2}{2(\beta\lambda)^{-1}}\right)\cdot\frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda} = \frac{b^a\sqrt{\beta}}{\Gamma(a)\sqrt{2\pi}}\lambda^{a-\frac{1}{2}}\exp\left(\left[-\frac{1}{2}(\mu-\mu_0)^2\beta-b\right]\lambda\right)$$

Given that we have $\mathbf{x} = \{x_1, \ldots, x_N\}$ i.i.d. observations, by the Bayes' formula, we obtain the relation

$$\begin{aligned}
p(\mu, \lambda | x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N | \mu, \lambda) p(\mu, \lambda) \\
&\propto \left[\prod_{i=1}^{N} p(x_i \mid \lambda)\right] p(\mu, \lambda) \\
&\propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp\left(-\frac{(x_i - \mu)^2}{2\lambda^{-1}}\right)\right] \cdot \frac{1}{\sqrt{2\pi(\beta\lambda)^{-1}}} \exp\left(-\frac{(\mu - \mu_0)^2}{2(\beta\lambda)^{-1}}\right) \cdot \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda} \\
&\propto \lambda^{\frac{N}{2}} \exp\left(\lambda \sum_{i=1}^{N} -\frac{1}{2}(x_i - \mu)^2\right) \cdot \sqrt{\beta\lambda} \exp\left(-\frac{(\mu - \mu_0)^2}{2(\beta\lambda)^{-1}}\right) \cdot \lambda^{a-1} e^{-b\lambda} \\
&\propto \lambda^{\frac{N}{2}+a-\frac{1}{2}} \exp\left(\lambda \cdot \left[-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\beta^{-1}} - b\right]\right) \\
&\propto \lambda^{\frac{N}{2}+a-\frac{1}{2}} \exp\left(\left[-\frac{1}{2}\left(\mu - \frac{\sum_{i=1}^{N} x_i + \beta\mu_0}{N+\beta}\right)^2 (N+\beta) + \frac{(\sum_{i=1}^{N} x_i + \beta\mu_0)^2}{2(N+\beta)} - \frac{1}{2}\sum_{i=1}^{N} x_i^2 - \frac{1}{2}\beta\mu_0^2 - b\right] \cdot \lambda\right) \\
\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \Longrightarrow &\propto \lambda^{\frac{N}{2}+a-\frac{1}{2}} \exp\left(\left[-\frac{1}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 (N+\beta) + \frac{(N\bar{x} + \beta\mu_0)^2}{2(N+\beta)} - \frac{1}{2}\left(\sum_{i=1}^{N}(x_i - \bar{x})^2 + N\bar{x}^2\right) - \frac{1}{2}\beta\mu_0^2 - b\right] \cdot \lambda\right) \\
&\propto \lambda^{\frac{N}{2}+a-\frac{1}{2}} \exp\left(\left[-\frac{1}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 (N+\beta) - \frac{1}{2}\frac{N\beta}{N+\beta}(\bar{x} - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{N}(x_i - \bar{x})^2 - b\right] \cdot \lambda\right)
\end{aligned}$$

Take $a' = \frac{N}{2} + a, \beta' = N + \beta, \mu_0' = \frac{N\bar{x}+\beta\mu_0}{N+\beta}$ and $b' = b + \frac{1}{2}\frac{N\beta}{N+\beta}(\bar{x} - \mu_0)^2 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \bar{x})^2$ where $\bar{x} = \sum_{i=1}^{N} x_i$, we attain

$$p(\mu, \lambda | x_1, \ldots, x_N) \propto \lambda^{a'-\frac{1}{2}} \exp\left(\left[-\frac{1}{2}(\mu - \mu_0')^2\beta' - b'\right]\lambda\right)$$

Hence, the posterior distribution is also a Gaussian-Gamma distribution, implying that the Gaussian-Gamma prior is a conjugate prior for the Gaussian likelihood function.

5 Topic: Bayesian Linear Regression and Feature Maps

Problem 5.1. *Let $\phi : \mathbb{R}^{1\times M} \to \mathbb{R}^{1\times M}, \mathbf{x} \mapsto \phi(\mathbf{x})$ be a feature map (also referred to as a feature expansion) where $\phi = (\phi_0, \ldots, \phi_{M-1})^\top$ and $\phi_i : \mathbb{R}^{1\times M} \to \mathbb{R}$.*

We consider the Bayesian linear regression problem

$$f(\mathbf{x}) := \phi(\mathbf{x})\mathbf{w}, \quad y = f(\mathbf{x}) + \varepsilon,$$

where $\mathbf{x} \in \mathbb{R}^{1\times M}$ is the feature vector as a row, $\mathbf{w} \in \mathbb{R}^M$ is a the unknown parameter vector, $y \in \mathbb{R}$ is the target, $\varepsilon \in \mathbb{R}$ is noise.

Given the training data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ where $\mathbf{X} \in \mathbb{R}^{N\times M}$, we define the design matrix by $\mathbf{\Phi}$, assume the residuals ε are i.i.d. Gaussian with mean zero and variance σ^2, and have the linear relation

$$\mathbf{\Phi} := \phi(\mathbf{X}) = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & & \vdots \\ \phi_0(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}, \quad \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right), \quad \mathbf{y} = \mathbf{\Phi}\mathbf{w} + \varepsilon,$$

respectively. Specifically, we choose a simple prior for $\mathbf{w}$ to be $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

(a) Following the exact same steps as in the derivation of Bayesian linear regression, the posterior distribution of $\mathbf{w}$ given the data $\mathcal{D}$ is

$$\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}, \mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}}\right)$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} &:= \mathbf{\Sigma}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y} \\
\mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} &= \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{\Phi}\mathbf{\Sigma}
\end{aligned}$$

Solution: Given $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ and $\mathbf{y} = \phi(\mathbf{X})\mathbf{w} + \varepsilon = \mathbf{\Phi}\mathbf{w} + \varepsilon$, by Item (e), we have

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})$$

Further, we can find the covariance between $\mathbf{y}$ and $\mathbf{w}$, we get

$$Cov(\mathbf{y}, \mathbf{w}) = \mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{w} - \mathbb{E}[\mathbf{w}])^\top\right] = \mathbb{E}\left[(\mathbf{\Phi}\mathbf{w} + \varepsilon - \mathbb{E}[\mathbf{\Phi}\mathbf{w} + \varepsilon])(\mathbf{w} - \mathbb{E}[\mathbf{w}])^\top\right] \xlongequal{\varepsilon \text{ independent}} \mathbf{\Phi}\mathbb{E}\left[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{w} - \mathbb{E}[\mathbf{w}])^\top\right] = \mathbf{\Phi}\mathbf{\Sigma}$$

By symmetry, we have $Cov(\mathbf{w}, \mathbf{y}) = \mathbf{\Sigma}\mathbf{\Phi}^\top$. Therefore, as in Problem 2.1, we can rewrite it into

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma}\mathbf{\Phi}^\top \\ \mathbf{\Phi}\mathbf{\Sigma} & \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I} \end{bmatrix}\right)$$

Directly applying Item (b) with $\mathbf{x}_1 = \mathbf{w}$, $\mathbf{x}_2 = \mathbf{y}$, $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{0}$, $\mathbf{\Sigma}_{11} = \mathbf{\Sigma}$, $\mathbf{\Sigma}_{12} = \mathbf{\Phi}\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{22} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}$, we obtain

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} &:= \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) = \mathbf{\Sigma}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y} \\
\mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} &:= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^\top = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}\right)^{-1}\mathbf{\Phi}\mathbf{\Sigma}
\end{aligned}$$

Note that in these formulas, the feature map ϕ appears in all of the expressions

$$\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top, \quad \mathbf{\Phi}_*\mathbf{\Sigma}\mathbf{\Phi}^\top, \quad \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}_*^\top, \quad \mathbf{\Phi}_*\mathbf{\Sigma}\mathbf{\Phi}_*^\top$$

The entries of these matrices are of the form

$$\phi(\mathbf{x})\mathbf{\Sigma}\phi(\mathbf{y})^\top = \phi(\mathbf{x})\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2^\top}\phi(\mathbf{y})^\top = \psi(\mathbf{x})\psi(\mathbf{y})^\top,$$

where $\mathbf{x}$ and $\mathbf{y}$ are two arbitrary inputs, $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2^\top}$ is the Cholesky decomposition of $\mathbf{\Sigma}$ as its positive definite, and $\psi(\mathbf{x}) = \phi(\mathbf{x})\mathbf{\Sigma}^{1/2}$.

To simply our expressions, we define the function

$$K(\mathbf{x}, \mathbf{y}) := \phi(\mathbf{x})\mathbf{\Sigma}\phi(\mathbf{y})^\top = \psi(\mathbf{x})\psi(\mathbf{y})^\top \text{ with } \mathbf{x}, \mathbf{y} \in \mathbb{R}^{1\times M} \quad \text{and} \quad K(\mathbf{X}, \mathbf{Y}) = \phi(\mathbf{X})\mathbf{\Sigma}\phi(\mathbf{Y})^\top = \begin{bmatrix} \phi(\mathbf{x}_1)\mathbf{\Sigma}\phi(\mathbf{y}_1)^\top & \cdots & \phi(\mathbf{x}_1)\mathbf{\Sigma}\phi(\mathbf{y}_N)^\top \\ \vdots & & \vdots \\ \phi(\mathbf{x}_N)\mathbf{\Sigma}\phi(\mathbf{y}_1)^\top & \cdots & \phi(\mathbf{x}_N)\mathbf{\Sigma}\phi(\mathbf{y}_N)^\top \end{bmatrix} \text{ with } \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times M}$$

Specifically, we have

$$K(\mathbf{X}, \mathbf{X}) = \phi(\mathbf{X})\mathbf{\Sigma}\phi(\mathbf{X})^\top = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top, \quad K(\mathbf{X}_*, \mathbf{X}) = \phi(\mathbf{X}_*)\mathbf{\Sigma}\phi(\mathbf{X})^\top = \mathbf{\Phi}_*\mathbf{\Sigma}\mathbf{\Phi}^\top, \quad K(\mathbf{X}, \mathbf{X}_*) = \phi(\mathbf{X})\mathbf{\Sigma}\phi(\mathbf{X}_*)^\top = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}_*^\top, \quad K(\mathbf{X}_*, \mathbf{X}_*) = \phi(\mathbf{X}_*)\mathbf{\Sigma}\phi(\mathbf{X}_*)^\top = \mathbf{\Phi}_*\mathbf{\Sigma}\mathbf{\Phi}_*^\top \tag{1}$$

(b) Using the posterior distribution for $\mathbf{w}$ *and the relation* $\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \varepsilon$*, we can predict the distribution for* $\mathbf{y}_*$ *given new data points* $\mathbf{X}_*$ *by*

$$\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2 \sim \mathcal{N}\left(\mathbf{\Phi}_* \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}, \quad \mathbf{\Phi}_* \mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} \mathbf{\Phi}_*^\top + \sigma^2 \mathbf{I}\right) = \mathcal{N}\left(\mu_{\mathbf{y}_*|\mathcal{D}}, K_{\mathbf{y}_*|\mathcal{D}} + \sigma^2 \mathbf{I}\right),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{y}_*|\mathcal{D}} &= K\left(\mathbf{X}_*, \mathbf{X}\right)\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y} \\
K_{\mathbf{y}_*|\mathcal{D}} &= K\left(\mathbf{X}_*, \mathbf{X}_*\right) - K\left(\mathbf{X}_*, \mathbf{X}\right)\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right)^{-1} K\left(\mathbf{X}, \mathbf{X}_*\right)
\end{aligned}$$

Solution: From the relation above and Equation (1), we get

$$\mu_{\mathbf{y}_*|\mathcal{D}} = \mathbf{\Phi}_* \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} = \mathbf{\Phi}_* \mathbf{\Sigma} \mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y} = K\left(\mathbf{X}_*, \mathbf{X}\right)\left(K\left(\mathbf{X}, \mathbf{X}\right) + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

and

$$\begin{aligned}
K_{\mathbf{y}_*|\mathcal{D}} &= \mathbf{\Phi}_* \mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} \mathbf{\Phi}_*^\top + \sigma^2 \mathbf{I} - \sigma^2 \mathbf{I} \\
&= \mathbf{\Phi}_* \left(\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{\Phi}\mathbf{\Sigma}\right) \mathbf{\Phi}_*^\top \\
&= \mathbf{\Phi}_* \mathbf{\Sigma} \mathbf{\Phi}_*^\top - \mathbf{\Phi}_* \mathbf{\Sigma} \mathbf{\Phi}^\top \left(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}_*^\top \\
&= K\left(\mathbf{X}_*, \mathbf{X}_*\right) - K\left(\mathbf{X}_*, \mathbf{X}\right)\left(K\left(\mathbf{X}, \mathbf{X}\right) + \sigma^2 \mathbf{I}\right)^{-1} K\left(\mathbf{X}, \mathbf{X}_*\right)
\end{aligned}$$