Solutions Manual to
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

January 30, 2025

Contents

1 Introduction

1.1

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, \mathbf{w}) - t_n \right)^2 . \tag{1.1}$$

Setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N} \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} \left(y(x_n, \mathbf{w}) - t_n \right) . \tag{1.2}$$

If

$$y(x_n, \mathbf{w}) = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(x_n), \tag{1.3}$$

then

$$\mathbf{0} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \left(\mathbf{w}^\mathsf{T} \boldsymbol{\phi}(x_n) - t_n \right) . \tag{1.4}$$

Therefore,

$$\left(\sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\mathsf{T} \right) \mathbf{w} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.5}$$

Thus,

$$\operatorname*{argmin}_{\mathbf{w}} E(\mathbf{w}) = \mathbf{A}^{-1} \mathbf{v}, \tag{1.6}$$

where

$$\mathbf{A} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\mathsf{T},$$
$$\mathbf{v} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.7}$$

If

$$\boldsymbol{\phi}(x_n) = \begin{bmatrix} 1 \\ x_n \\ \vdots \\ x_n^M \end{bmatrix},$$

1

then

$$A_{mm'} = \sum_{n=1}^{N} x_n^{m+m'},$$

$$v_m = \sum_{n=1}^{N} t_n x_n^m. \tag{1.8}$$

1.2

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2. \tag{1.9}$$

Setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N} \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} (y(x_n, \mathbf{w}) - t_n) + \lambda \mathbf{w}. \tag{1.10}$$

If

$$y(x_n, \mathbf{w}) = \mathbf{w}^\intercal \boldsymbol{\phi}(x_n), \tag{1.11}$$

then

$$\mathbf{0} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \left(\mathbf{w}^\intercal \boldsymbol{\phi}(x_n) - t_n\right) + \lambda \mathbf{w}. \tag{1.12}$$

Therefore,

$$\left(\sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^\intercal + \lambda \mathbf{I}\right) \mathbf{w} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.13}$$

Thus,

$$\underset{\mathbf{w}}{\operatorname{argmin}} \, E(\mathbf{w}) = \mathbf{A}^{-1}\mathbf{v}, \tag{1.14}$$

where

$$\mathbf{A} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^\intercal + \lambda \mathbf{I},$$

$$\mathbf{v} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.15}$$

If

$$\phi(x_n) = \begin{bmatrix} 1 \\ x_n \\ \vdots \\ x_n^M \end{bmatrix},$$

then

$$A_{mm'} = \sum_{n=1}^{N} x_n^{m+m'} + \lambda I_{mm'},$$

$$v_m = \sum_{n=1}^{N} t_n x_n^m. \tag{1.16}$$

1.3

Let a, o and l be the events where an apple, orange and lime are selected respectively. The probability that an apple is selected is given by

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \tag{1.17}$$

Substituting $p(a|r) = \frac{3}{10}$, $p(r) = \frac{1}{5}$, $p(a|g) = \frac{1}{2}$, $p(r) = \frac{1}{5}$, $p(a|g) = \frac{3}{10}$ and $p(g) = \frac{3}{5}$ gives

$$p(a) = \frac{17}{50}. \tag{1.18}$$

If an orange is selected, the probability that it came from the geen box is given by

$$p(g|o) = \frac{p(g, o)}{p(o)}. \tag{1.19}$$

Here,

$$p(g, o) = p(o|g)p(g),$$
$$p(o) = p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g). \tag{1.20}$$

Substituting $p(o|r) = \frac{2}{5}$, $p(r) = \frac{1}{5}$, $p(o|b) = \frac{1}{2}$, $p(b) = \frac{1}{5}$, $p(o|g) = \frac{3}{10}$ and $p(g) = \frac{3}{5}$ gives $p(g, o) = \frac{9}{50}$ and $p(o) = \frac{9}{25}$. Therefore,

$$p(g|o) = \frac{1}{2}. \tag{1.21}$$

1.4

Let

$$x = g(y) \tag{1.22}$$

and $\hat{x}$ and $\hat{y}$ be the locations of the maximum of $p_x(x)$ and $p_y(y)$ respcetively. Let us assume that there exists $\epsilon > 0$ such that $g'(y) \neq 0$ for $|y - \hat{y}| < \epsilon$. Then, Taking the derivative of the transoformation

$$p_y(y) = p_x\left(g(y)\right)\left|g'(y)\right| \tag{1.23}$$

and substituting $y = \hat{y}$ gives

$$0 = g'(\hat{y})p'_x\left(g\left(\hat{y}\right)\right) + p_x\left(g\left(\hat{y}\right)\right)g''\left(\hat{y}\right). \tag{1.24}$$

Therefore, in general,

$$\hat{x} \neq g\left(\hat{y}\right). \tag{1.25}$$

Here, let us assume that

$$g(y) = ay + b. \tag{1.26}$$

Then, Taking the derivative of the transformation and substituting $y = \hat{y}$ gives

$$0 = p'_x\left(g\left(\hat{y}\right)\right). \tag{1.27}$$

Therefore,

$$\hat{x} = g\left(\hat{y}\right). \tag{1.28}$$

1.5

By the definition,

$$\operatorname{var} f(x) = \operatorname{E}\left(f(x) - \operatorname{E} f(x)\right)^2. \tag{1.29}$$

The right hand side can be written as

$$\operatorname{E}\left((f(x))^2 - 2f(x)\operatorname{E} f(x) + (\operatorname{E} f(x))^2\right) = \operatorname{E}\left(f(x)\right)^2 - (\operatorname{E} f(x))^2. \tag{1.30}$$

Therefore,

$$\operatorname{var} f(x) = \operatorname{E}\left(f(x)\right)^2 - (\operatorname{E} f(x))^2. \tag{1.31}$$

1.6

By the definition,

$$\text{cov}(x, y) = \text{E}\left((x - \text{E}\,x)(y - \text{E}\,y)\right). \tag{1.32}$$

The right hand side can be written as

$$\text{E}\,xy - \text{E}\,(x\,\text{E}\,y) - \text{E}\,(y\,\text{E}\,x) + \text{E}\,(\text{E}\,x\,\text{E}\,y) = \text{E}\,xy - \text{E}\,x\,\text{E}\,y. \tag{1.33}$$

The right hand side can be written as

$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \tag{1.34}$$

If x and y are independent, by the definition,

$$f(x, y) = f(x)f(y). \tag{1.35}$$

Then,

$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \tag{1.36}$$

Therefore,

$$\text{cov}(x, y) = 0. \tag{1.37}$$

1.7

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \tag{1.38}$$

Then

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left(x^2 + y^2\right)\right) dxdy. \tag{1.39}$$

By the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ), the right hand side can be written as

$$\int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} drd\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.40}$$

5

By the transformation $s = \frac{r}{\sigma}$, the right hand side can be written as

$$2\pi\sigma^2 \int_0^\infty \exp\left(-\frac{1}{2}s^2\right) s\,ds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^\infty. \tag{1.41}$$

Therefore,

$$I = \left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.42}$$

By the definition,

$$\mathcal{N}\left(x|\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.43}$$

Then

$$\int_{-\infty}^\infty \mathcal{N}\left(x|\mu, \sigma^2\right) dx = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \tag{1.44}$$

By the transformation $t = x - \mu$, the right hand side can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} I. \tag{1.45}$$

Therefore,

$$\int_{-\infty}^\infty \mathcal{N}\left(x|\mu, \sigma^2\right) dx = 1. \tag{1.46}$$

1.8

Let x be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \tag{1.47}$$

Then

$$\mathbb{E}\,x = \int_{-\infty}^\infty x\mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.48}$$

By the definition, the right hand side can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^\infty x\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \tag{1.49}$$

By the transformation $y = x - \mu$, it can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y + \mu) \exp\left(-\frac{1}{2\sigma^2} y^2\right) dy. \tag{1.50}$$

Since

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2} y^2\right) dy = 0, \tag{1.51}$$

and

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2} y^2\right) dy = \mu \int_{-\infty}^{\infty} \mathcal{N}\left(y|\mu, \sigma^2\right) dy, \tag{1.52}$$

we have

$$\mathrm{E}\, x = \mu. \tag{1.53}$$

By the definition,

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx = 1 \tag{1.54}$$

can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx = 1. \tag{1.55}$$

Taking the derivative with respect to σ^2 gives

$$(2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) \left(\sigma^2\right)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx$$
$$+ \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} \left(\sigma^2\right)^{-2} (x - \mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx = 0. \tag{1.56}$$

The left hand side can be written as

$$-\frac{1}{2} \left(\sigma^2\right)^{-1} \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx + \frac{1}{2} \left(\sigma^2\right)^{-2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}\left(x|\mu, \sigma^2\right) dx$$
$$= -\frac{1}{2} \left(\sigma^2\right)^{-1} + \frac{1}{2} \left(\sigma^2\right)^{-2} \operatorname{var} x. \tag{1.57}$$

Therefore,

$$\operatorname{var} x = \sigma^2. \tag{1.58}$$

1.9

Let

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.59}$$

Setting its derivative with respect to x to zero gives

$$0 = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x-\mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.60}$$

Therefore, the mode is given by μ.

Similarly, let

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = (2\pi)^{-\frac{D}{2}} \left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.61}$$

Setting its derivative with respect to $\mathbf{x}$ to zero gives

$$\mathbf{0} = -\left(2\pi\right)^{-\frac{D}{2}} \left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right) (\mathbf{x}-\boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.62}$$

Therefore, the mode is given by $\boldsymbol{\mu}$.

1.10

By the definition,

$$\mathrm{E}(x+y) = \int\int (x+y)p(x,y)dxdy. \tag{1.63}$$

The right hand side can be written as

$$\int x\left(\int p(x,y)dy\right) dx + \int y\left(\int p(x,y)dx\right) dy = \int xp(x)dx + \int yp(y)dy. \tag{1.64}$$

By the definition, the right hand side can be written as

$$\mathrm{E}\,x + \mathrm{E}\,y. \tag{1.65}$$

Therefore,

$$\mathrm{E}(x+y) = \mathrm{E}\,x + \mathrm{E}\,y. \tag{1.66}$$

Similarly, by the definition,

$$\text{var}(x + y) = \text{E}\left(x + y - \text{E}(x + y)\right)^2 \tag{1.67}$$

By the result above and the definition, the right hand side can be written as

$$\text{E}\left(x - \text{E}\,x\right)^2 + 2\,\text{E}\left((x - \text{E}\,x)(y - \text{E}\,y)\right) + \text{E}\left(y - \text{E}\,y\right)^2$$
$$= \text{var}\,x + 2\,\text{cov}(x, y) + \text{var}\,y. \tag{1.68}$$

If x and y are independent, then

$$\text{cov}(x, y) = 0, \tag{1.69}$$

by 1.6. Therefore,

$$\text{var}(x + y) = \text{var}\,x + \text{var}\,y. \tag{1.70}$$

1.11

Let $x_1, \cdots, x_N$ be variables such that

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right). \tag{1.71}$$

Then

$$\ln p(\mathbf{x}) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.72}$$

To maximise it with respect to μ and σ^2, setting the partial derivatives to zero gives

$$0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu),$$
$$0 = -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.73}$$

Therefore,

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n,$$
$$\sigma_{\text{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2. \tag{1.74}$$

9

1.12

Let x_m and x_n be independent variables. Then

$$\mathrm{E}\,x_m x_n = \mathrm{E}\,x_m\,\mathrm{E}\,x_n. \tag{1.75}$$

If they are samples from the Gaussian distribution with mean μ and variance σ^2, the right hand side is given by μ^2. On the other hand, by the definition,

$$\mathrm{E}\,x_n^2 = \mathrm{var}\,x_n + \left(\mathrm{E}\,x_n\right)^2. \tag{1.76}$$

If x_n is a sample from the Gaussian distribution with mean μ and variance σ^2, the right hand side is given by $\sigma^2 + \mu^2$. Therefore,

$$\mathrm{E}\,x_m x_n = \mu^2 + \delta_{mn}\sigma^2. \tag{1.77}$$

Here, since

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{1.78}$$

we have

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\,x_n. \tag{1.79}$$

Therefore,

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \mu. \tag{1.80}$$

Similarly, since

$$\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \mu_{\mathrm{ML}}\right)^2, \tag{1.81}$$

we have

$$\mathrm{E}\,\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.82}$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n^2 - 2\mu_{\mathrm{ML}}x_n + \mu_{\mathrm{ML}}^2\right) = \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\,x_n^2 - \frac{2}{N}\,\mathrm{E}\left(\mu_{\mathrm{ML}}\left(\sum_{n=1}^{N} x_n\right)\right) + \mathrm{E}\,\mu_{\mathrm{ML}}^2. \tag{1.83}$$

The first term of the right hand side can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.84}$$

while the second and third terms can be writen as

$$-2\,\mathrm{E}\,\mu_{\mathrm{ML}}^2+\mathrm{E}\,\mu_{\mathrm{ML}}^2=-\,\mathrm{E}\,\mu_{\mathrm{ML}}^2. \tag{1.85}$$

Here,

$$\mathrm{E}\,\mu_{\mathrm{ML}}^2=\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2. \tag{1.86}$$

The right hand side can be written as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\,x_n^2+\frac{2}{N^2}\sum_{1\le m<n\le N}\mathrm{E}\,x_m x_n=\frac{1}{N}\left(\mu^2+\sigma^2\right)+\frac{N-1}{N}\mu^2. \tag{1.87}$$

Therefore,

$$\mathrm{E}\,\mu_{\mathrm{ML}}^2=\mu^2+\frac{1}{N}\sigma^2. \tag{1.88}$$

Thus,

$$\mathrm{E}\,\sigma_{\mathrm{ML}}^2=\frac{N-1}{N}\sigma^2. \tag{1.89}$$

1.13

Let $x_1,\cdots,x_N$ be a set of variables whose mean is μ and variance is σ^2. Then

$$\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\,(x_n-\mu)^2. \tag{1.90}$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n^2-2\mu x_n+\mu^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\,x_n^2-\frac{2\mu}{N}\sum_{n=1}^{N}\mathrm{E}\,x_n+\mu^2. \tag{1.91}$$

The first term of the right hand side can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.92}$$

while the second term can be writen as

$$-\frac{2\mu}{N}\sum_{n=1}^{N}\mu = -2\mu^2. \tag{1.93}$$

Therefore,

$$\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right) = \sigma^2. \tag{1.94}$$

1.14

Let

$$w_{ij}^{\mathrm{S}} = \frac{1}{2}(w_{ij} + w_{ji}),$$
$$w_{ij}^{\mathrm{A}} = \frac{1}{2}(w_{ij} - w_{ji}). \tag{1.95}$$

Then

$$w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}},$$
$$w_{ij}^{\mathrm{S}} = w_{ji}^{\mathrm{S}},$$
$$w_{ij}^{\mathrm{A}} = -w_{ji}^{\mathrm{A}}. \tag{1.96}$$

Here,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_i x_j = \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij} - w_{ji})x_i x_j. \tag{1.97}$$

The right hand side can be written as

$$\frac{1}{2}\left(\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_i x_j - \sum_{i=1}^{D}\sum_{j=1}^{D}w_{ji}x_i x_j\right) = 0. \tag{1.98}$$

Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_i x_j = 0. \tag{1.99}$$

Additionally,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D}\left(w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}\right)x_i x_j. \tag{1.100}$$

12

The right hand side can be written as

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\text{S}} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\text{A}} x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\text{S}} x_i x_j, \tag{1.101}$$

where the result above is used. Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\text{S}} x_i x_j. \tag{1.102}$$

Finally, since the matrix w_{ij}^{S} is a $D \times D$ symmetric matrix, its number of independent parameters is $\frac{D(D+1)}{2}$.

1.15

Let $n(D, M)$ be the number of independent parameters of a polynomial in D dimensions and M orders. Then

$$n(1, M) = n(1, M - 1) = 1. \tag{1.103}$$

Let us assume that

$$n(D, M) = \sum_{i=1}^{D} n(i, M - 1). \tag{1.104}$$

The independent terms of a polynomial in $D + 1$ dimensions and M orders can be split into 1. the ones of a polynomial in D dimensions and M orders and 2. the ones generated by multiplying the ones in $D + 1$ dimensions and M orders by the $D + 1$th variable. Therefore,

$$n(D + 1, M) = n(D, M) + n(D + 1, M - 1). \tag{1.105}$$

Thus,

$$n(D + 1, M) = \sum_{i=1}^{D+1} n(i, M - 1). \tag{1.106}$$

Hence, the assumption is proved by induction on D.

Additionally,

$$\sum_{i=1}^{1} \frac{(i + M - 2)!}{(i - 1)!(M - 1)!} = 1. \tag{1.107}$$

13

Let us assume that

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}. \tag{1.108}$$

Then

$$\sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!}. \tag{1.109}$$

The right hand side can be written as

$$\frac{D(D+M-1)! + M(D+M-1)!}{D!M!} = \frac{(D+M)!}{D!M!}. \tag{1.110}$$

Therefore, the assumption is proved by induction on D.

Finally, by 1.14,

$$n(D, 2) = \frac{D(D+1)}{2}. \tag{1.111}$$

Let us assume that

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!}. \tag{1.112}$$

Then, by the result above,

$$n(D, M+1) = \sum_{i=1}^{D} n(i, M). \tag{1.113}$$

By the assumption and result above, the right hand side can be written as

$$\sum_{i=1}^{D} \frac{(i+M-1)!}{(i-1)!M!} = \frac{(D+M)!}{(D-1)!(M+1)!}. \tag{1.114}$$

Therefore, the assumption is proved by induction on M.

1.16

Let $N(D, M)$ be the number of independent parameters in all of the terms up to and including the ones of D dimensions and M orders. Then, by 1.15,

$$N(D, M) = \sum_{m=0}^{M} n(D, m), \tag{1.115}$$

where
$$n(D, m) = \frac{(D + m - 1)!}{(D - 1)!m!}. \tag{1.116}$$

Additionally,
$$N(D, 0) = 1. \tag{1.117}$$

Let us assume that
$$\sum_{m=0}^{M} n(D, m) = \frac{(D + M)!}{D!M!}. \tag{1.118}$$

Then
$$\sum_{m=0}^{M+1} n(D, m) = \frac{(D + M)!}{D!M!} + \frac{(D + M)!}{(D - 1)!(M + 1)!}. \tag{1.119}$$

The right hand side can be written as
$$\frac{(M + 1)(D + M)! + D(D + M)!}{D!(M + 1)!} = \frac{(D + M + 1)!}{D!(M + 1)!}. \tag{1.120}$$

Therefore, the assumption is proved by induction on M. Thus,
$$N(D, M) = \frac{(D + M)!}{D!M!}. \tag{1.121}$$

Additionally, by the approximation
$$n! \simeq n^n \exp(-n), \tag{1.122}$$

the right hand side can be approximated as
$$\frac{(D + M)^{D+M} \exp\left(-(D + M)\right)}{D^D \exp(-D) M^M \exp(-M)} = \frac{(D + M)^{D+M}}{D^D M^M}. \tag{1.123}$$

The right hand side can be written as
$$D^M \left(1 + \frac{M}{D}\right)^D \left(\frac{1}{M} + \frac{1}{D}\right)^M = M^D \left(1 + \frac{D}{M}\right)^M \left(\frac{1}{D} + \frac{1}{M}\right)^D. \tag{1.124}$$

Therefore, $N(D, M)$ can be approximated as D^M for $D \gg M$ and as M^D for $M \gg D$.

Finally, by the result above,
$$\begin{aligned} N(10, 3) &= 286, \\ N(100, 3) &= 176851, \\ N(1000, 3) &= 167668501. \end{aligned} \tag{1.125}$$

1.17

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \tag{1.126}$$

Then

$$\Gamma(x+1) = \int_0^\infty u^x \exp(-u) du. \tag{1.127}$$

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x\Gamma(x). \tag{1.128}$$

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.129}$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u) du, \tag{1.130}$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \tag{1.131}$$

For a positive integer x, let us assume that

$$\Gamma(x) = (x-1)!. \tag{1.132}$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \tag{1.133}$$

where the right hand side can be written as $x!$. Therefore,

$$\Gamma(x+1) = x!. \tag{1.134}$$

Thus, the assumption is proved by induction on x.

1.18

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} \exp\left(-x_i^2\right) dx_i = S_D \int_0^{\infty} \exp\left(-r^2\right) r^{D-1} dr, \tag{1.135}$$

where S_D is the surface area of a sphere of unit raidus in D dimensions. By 1.7, the left hand side can be written as $\pi^{\frac{D}{2}}$. By the transformation $s = r^2$, the right hand side can be written as

$$\frac{S_D}{2} \int_0^{\infty} \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \tag{1.136}$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \tag{1.137}$$

Additionally, the volume of the sphere can can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \tag{1.138}$$

The right hand side can be written as

$$S_D \left[\frac{r^D}{D}\right]_{r=0}^{r=1} = \frac{S_D}{D}. \tag{1.139}$$

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.140}$$

Finally, the results above reduce to

$$S_2 = \frac{2\pi}{\Gamma(1)},$$
$$V_2 = \frac{S_2}{2}. \tag{1.141}$$

Therefore,

$$S_2 = 2\pi,$$
$$V_2 = \pi. \tag{1.142}$$

17

Similarly,

$$S_3 = \frac{2\pi^{\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)},$$

$$V_3 = \frac{S_3}{3}.$$

(1.143)

Therefore,

$$S_3 = 4\pi,$$

$$V_3 = \frac{4}{3}\pi.$$

(1.144)

1.19

The volume of a cube of side 2 in D dimensions is 2^D. Therefore, the ratio of the volume of the cocentric sphere of radius 1 divided by the volume of the cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma\left(\frac{D}{2}\right)},$$

(1.145)

by 1.18.

Additionally, by Stering's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}}\exp(-x)x^{\frac{x+1}{2}},$$

(1.146)

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}}\exp\left(1-\frac{D}{2}\right)\left(\frac{D}{2}-1\right)^{\frac{D}{4}}}.$$

(1.147)

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}}\frac{1}{D}\left(\frac{e^2\pi^2}{8D-16}\right)^{\frac{D}{4}}.$$

(1.148)

Therefore, the ratio goes to zero as $D \to \infty$.

Finally, the ratio of the distance from the center of the cube to one of the corners divided by the perpendicular distance to one of the sides is given by

$$\frac{\sqrt{\sum_{i=1}^{D}1^2}}{1} = \sqrt{D}.$$

(1.149)

Therefore, the ration goes to ∞ as $D \to \infty$.

1.20

For a vector $\mathbf{x}$ in D dimensions, let

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \tag{1.150}$$

Then

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_r^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\boldsymbol{\phi}, \tag{1.151}$$

where $\boldsymbol{\phi}$ is the vector of the angular components of the polar corrdinate and J is the Jacobian of the transformation from the Cartesian to polar coordinate. For a sufficiently small ϵ, the right hand side can be approximated as

$$(2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\boldsymbol{\phi}$$

$$= (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}. \tag{1.152}$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r)\epsilon, \tag{1.153}$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{1.154}$$

and S_D is the surface area of a unit sphere in D dimensions.

Additionally, setting the derivative of $p(r)$ to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left((D-1)r^{D-2} - \frac{r^D}{\sigma^2}\right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{1.155}$$

Therefore, $p(r)$ is maximised at a sigle stationary point

$$\hat{r} = \sqrt{D-1}\sigma. \tag{1.156}$$

Additionally, by the expression of $p(r)$ above,

$$\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} = \left(\frac{\hat{r}+\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \tag{1.157}$$

Using the expression of $\hat{r}$ above, the right hand side can be written as

$$\exp\left((D-1)\ln\left(1+\frac{\epsilon}{\hat{r}}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right)$$
$$=\exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1+\frac{\epsilon}{\hat{r}}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right). \tag{1.158}$$

By the Taylor series

$$\ln(1+x)=x-\frac{1}{2}x^2+o\left(x^3\right), \tag{1.159}$$

the right hand side can be approximated as

$$\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}}-\frac{\epsilon^2}{2\hat{r}^2}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right)=\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.160}$$

Therefore,

$$p\left(\hat{r}+\epsilon\right)\simeq p\left(\hat{r}\right)\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.161}$$

Finally, let a vector of length $\hat{r}$ be $\hat{\mathbf{r}}$. Then, by the definition of $p(\mathbf{x})$,

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)}=\exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \tag{1.162}$$

Substituting the expression of $\hat{r}$ above, the right hand side can be written as $\exp\left(\frac{D-1}{2}\right)$. Therefore,

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)}=\exp\left(\frac{D-1}{2}\right). \tag{1.163}$$

1.21

If $0\leq a\leq b$, then

$$0\leq a(b-a). \tag{1.164}$$

Therefore,

$$a\leq(ab)^{\frac{1}{2}}. \tag{1.165}$$

For a two-class classification problem of $\mathbf{x}$, let the classes be $\mathcal{C}_1$ and $\mathcal{C}_2$ and let the decision regions be $\mathcal{R}_1$ and $\mathcal{R}_2$. Let us choose the decision regions to minimise the probability of misclassification. Then,

$$p(\mathbf{x},\mathcal{C}_1)>p(\mathbf{x},\mathcal{C}_2)\Rightarrow\mathbf{x}\in\mathcal{C}_1, \tag{1.166}$$

20

and

$$p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1) \Rightarrow \mathbf{x} \in \mathcal{C}_2. \tag{1.167}$$

Then, using the inequality above,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} \le \int_{\mathcal{R}_1} \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2) \right)^{\frac{1}{2}} d\mathbf{x}, \tag{1.168}$$

and

$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \le \int_{\mathcal{R}_2} \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2) \right)^{\frac{1}{2}} d\mathbf{x}. \tag{1.169}$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \le \int \left(p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2) \right)^{\frac{1}{2}} d\mathbf{x}. \tag{1.170}$$

1.22

Let

$$\mathrm{E}\, L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.171}$$

If

$$L_{kj} = 1 - \delta_{kj}, \tag{1.172}$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \tag{1.173}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \tag{1.174}$$

Therefore,

$$\mathrm{E}\, L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.175}$$

Thus, minimising $\mathrm{E}\, L$ reduces to choosing the criterion to maximise the posterior probatility $p(\mathcal{C}_j|\mathbf{x})$.

1.23

Let

$$\mathrm{E}\,L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \qquad (1.176)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \qquad (1.177)$$

Therefore,

$$\mathrm{E}\,L = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \qquad (1.178)$$

Thus, minimising $\mathrm{E}\,L$ reduces to choosing to minimise $\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$.

1.24 (Incomplete)

Let

$$\mathrm{E}\,L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} + \lambda \int_{\forall k p(\mathcal{C}_k | \mathbf{x}) < \theta} p(\mathbf{x}) d\mathbf{x}. \qquad (1.179)$$

1.25

Let

$$\mathrm{E}\,L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \qquad (1.180)$$

Setting the derivative with respect to $\mathbf{y}(\mathbf{x})$ to zero gives

$$\mathbf{0} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t})\, p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \qquad (1.181)$$

The integral of the right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) p(\mathbf{x}) - p(\mathbf{x}) \int \mathbf{t} p(\mathbf{t} | \mathbf{x}) d\mathbf{t}. \qquad (1.182)$$

The integral in the second term of the right hand side can be written as $\mathrm{E}_{\mathbf{t}}(\mathbf{t} | \mathbf{x})$. Therefore, the right hand side can be written as

$$\mathbf{0} = p(\mathbf{x}) \left(\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t} | \mathbf{x}) \right). \qquad (1.183)$$

Thus,
$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} \operatorname{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.184}$$

Finally, for a single target variable t, it reduces to
$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} \operatorname{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \operatorname{E}_t(t|\mathbf{x}). \tag{1.185}$$

1.26

Let
$$\operatorname{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.186}$$

The right hand side can be written as
$$\int \int \|\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}$$
$$= \int \int \|\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}$$
$$+ 2 \int \int (\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^{\mathsf{T}} (\operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}$$
$$+ \int \int \|\operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.187}$$

Let us look at each term of the right hand side. The first term can be written as
$$\int \|\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left(\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \tag{1.188}$$

The second term can be written as
$$2 \int (\mathbf{y}(\mathbf{x}) - \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^{\mathsf{T}} \left(\int (\operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.189}$$

Since
$$\int \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t}}{p(\mathbf{x})},$$
$$\int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \operatorname{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}), \tag{1.190}$$

the second term is zero. The third term can be written as

$$\int \left(\int \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x})d\mathbf{t} \right) p(\mathbf{x})d\mathbf{x} = \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \qquad (1.191)$$

Therefore,

$$\mathrm{E}\, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x})d\mathbf{x} + \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \qquad (1.192)$$

Thus,

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} \mathrm{E}\, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \qquad (1.193)$$

1.27 (Incomplete)

Let

$$\mathrm{E}\, L_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \qquad (1.194)$$

Setting the derivative with respect to $y(\mathbf{x})$ to zero gives

$$0 = q p(\mathbf{x}) \int |y(\mathbf{x}) - t|^{q-1} \mathrm{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \qquad (1.195)$$

Therefore,

$$\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_q = \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \mathrm{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = 0 \right\}. \qquad (1.196)$$

Additionally, if $q = 1$, the integral can be written as

$$p(\mathbf{x}) \int \mathrm{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = p(\mathbf{x}) \left(\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x}) dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x}) dt \right). \qquad (1.197)$$

Therefore,

$$\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_1 = \mathrm{median}(t|\mathbf{x}). \qquad (1.198)$$

Finally,

$$\lim_{q \to 0} \left(\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_q \right) = \mathrm{mode}(t|\mathbf{x})? \qquad (1.199)$$

1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \tag{1.200}$$

Let $h(p)$ be a function to relate h and p. Then

$$h\left(p^2\right) = 2h(p). \tag{1.201}$$

Let us assume that, for a positive integer n,

$$h\left(p^n\right) = nh(p). \tag{1.202}$$

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p). \tag{1.203}$$

Therefore,

$$h\left(p^{n+1}\right) = (n+1)h(p). \tag{1.204}$$

Thus, the second assumption is proved by induction on n.

Additionally, for positive integers m and n,

$$h\left(p^n\right) = h\left(p^{\frac{n}{m}m}\right). \tag{1.205}$$

By the second assumption, the left hand side can be written as $nh(p)$. By the first assumption, the right hand side can be written as $mh\left(p^{\frac{n}{m}}\right)$. Therefore,

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \tag{1.206}$$

Finally, by the continuity, for a positive real number a,

$$h\left(p^a\right) = ah(p). \tag{1.207}$$

Taking the derivative with respect to a and substituting $a = 1$ gives

$$(p \ln p)h'(p) = h(p). \tag{1.208}$$

Therefore,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + \text{const}. \tag{1.209}$$

Ignoring the constants, the left hand side can be written as $\ln h(p)$ and the right hand side can be written as $\ln(\ln p)$. Thus,

$$h(p) \propto \ln p. \tag{1.210}$$

1.29

Let x be an M-state discrete random variable. Then, by the definition,

$$\mathrm{H}(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i), \tag{1.211}$$

where

$$\sum_{i=1}^{M} p(x_i) = 1. \tag{1.212}$$

By Jensen's inequality,

$$\sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \le \ln \left(\sum_{i=1}^{M} 1 \right). \tag{1.213}$$

Therefore,

$$\mathrm{H}(x) \le \ln M. \tag{1.214}$$

1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}\left(x|\mu, \sigma^2\right), \\ q(x) &= \mathcal{N}\left(x|m, s^2\right). \end{aligned} \tag{1.215}$$

By the definition,

$$\mathrm{KL}\left(p||q\right) = -\int p(x) \ln \frac{q(x)}{p(x)} dx. \tag{1.216}$$

The right hand side can be written as

$$\begin{aligned} &-\int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\ &= -\int_{-\infty}^{\infty} p(x) \left(-\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx. \end{aligned} \tag{1.217}$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \tag{1.218}$$

The first term can be written as $\ln \frac{s}{\sigma}$. The second term can be written as

$$\frac{1}{2s^2} \int_{-\infty}^{\infty} (x - \mu + \mu - m)^2 p(x) dx = \frac{\sigma^2 + (\mu - m)^2}{2s^2}. \tag{1.219}$$

The third term can be written as $-\frac{1}{2}$. Therefore,

$$\mathrm{KL}\,(p||q) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2}. \tag{1.220}$$

1.31

Let $\mathbf{x}$ and $\mathbf{y}$ be two variables. Then, by the definition,

$$\mathrm{H}(\mathbf{x}) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x},$$

$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \tag{1.221}$$

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = -\int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

Note that

$$\mathrm{H}(\mathbf{x}) = -\int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x},$$

$$\mathrm{H}(\mathbf{y}) = -\int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} \right) \ln p(\mathbf{y}) d\mathbf{y}. \tag{1.222}$$

Therefore,

$$\mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{x}, \mathbf{y}) = -\int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \tag{1.223}$$

Since

$$\int \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1, \tag{1.224}$$

Jensen's inequality can be used to write that

$$-\int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y} \geq -\ln \left(\int \int p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \right). \tag{1.225}$$

The right hand side can be written as

$$-\ln \left(\int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \tag{1.226}$$

Thus,

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) \leq \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}). \tag{1.227}$$

1.32

Let $\mathbf{x}$ be a vector of continuous variables and

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad (1.228)$$

where $\mathbf{A}$ is a nonsingular matrix. By the definition,

$$H(\mathbf{y}) = -\int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \qquad (1.229)$$

By the transformation

$$p_y(\mathbf{y}) = p_x(\mathbf{A}\mathbf{x}) |\det \mathbf{A}^{-1}|, \qquad (1.230)$$

the right hand side can be written as

$$-\int p_x(\mathbf{A}\mathbf{x}) \ln p_x(\mathbf{A}\mathbf{x}) |\det \mathbf{A}| d\mathbf{x} - \ln |\det \mathbf{A}^{-1}| \int p_y(\mathbf{y}) d\mathbf{y}. \qquad (1.231)$$

By the transformation

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \qquad (1.232)$$

the first term can be written as

$$-\int p_x(\mathbf{x}') \ln p_x(\mathbf{x}') d\mathbf{x}' = H(\mathbf{x}), \qquad (1.233)$$

and the second term can be written as

$$-\ln |\det \mathbf{A}^{-1}| = \ln |\det \mathbf{A}|. \qquad (1.234)$$

Therefore,

$$H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{A}|. \qquad (1.235)$$

1.33

Let x and y be two discrete variables. By the definition,

$$H(y|x) = -\sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \qquad (1.236)$$

28

If $H(y|x)$ is zero, then

$$0 = -\sum_i p(x_i) \sum_j p(y_j|x_i) \ln p(y_j|x_i). \qquad (1.237)$$

Since

$$p(x_i) \geq 0,$$
$$p(y_j|x_i) \ln p(y_j|x_i) \leq 0. \qquad (1.238)$$

for all i and j, the equation reduces to

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \qquad (1.239)$$

Therefore, $p(y_j|x_i)$ is zero or one. Thus, since

$$\sum_j p(y_j|x_i) = 1, \qquad (1.240)$$

$p(y_j|x_i)$ is one for a unique x_i and zero for others.

1.34

Let

$$L(p(x)) = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right)$$
$$+ \lambda_2 \left(\int_{-\infty}^{\infty} x p(x) dx - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right). \qquad (1.241)$$

Setting the derivtive with respect to $p(x)$ to zero gives

$$0 = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2. \qquad (1.242)$$

Therefore,
$$p(x) = \exp\left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2\right). \qquad (1.243)$$

Therefore,

$$p(x) = \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} + \lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right). \qquad (1.244)$$

Substituting it to

$$\int_{-\infty}^{\infty} p(x)dx = 1,$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu, \tag{1.245}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2,$$

gives

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx = 1,$$

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} x \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx = \mu,$$

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx = \sigma^2.$$

$$\tag{1.246}$$

By the transformation

$$y = \sqrt{-\lambda_3} \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right), \tag{1.247}$$

they can be written as

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = 1,$$

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}}y + \mu - \frac{\lambda_2}{2\lambda_3}\right) \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = \mu,$$

$$\exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}}y - \frac{\lambda_2}{2\lambda_3}\right)^2 \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = \sigma^2.$$

$$\tag{1.248}$$

Since

$$\int_{-\infty}^{\infty} \exp\left(-y^2\right) dy = \Gamma\left(\frac{1}{2}\right),$$

$$\int_{-\infty}^{\infty} y \exp\left(-y^2\right) dy = 0, \tag{1.249}$$

$$\int_{-\infty}^{\infty} y^2 \exp\left(-y^2\right) dy = \Gamma\left(\frac{3}{2}\right),$$

they can be written as

$$\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)=1,$$

$$\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)=\mu,$$

$$\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\left((-\lambda_3)^{-\frac{3}{2}}\Gamma\left(\frac{3}{2}\right)+(-\lambda_3)^{-\frac{1}{2}}\frac{\lambda_2^2}{4\lambda_3^2}\Gamma\left(\frac{1}{2}\right)\right)=\sigma^2.$$

$$(1.250)$$

Therefore,

$$\lambda_1 = 1-\frac{1}{2}\ln\left(2\pi\sigma^2\right),$$

$$\lambda_2 = 0, \tag{1.251}$$

$$\lambda_3 = -\frac{1}{2\sigma^2}.$$

Thus,

$$p(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.252}$$

1.35

Let x be a variable such that

$$p(x) = \mathcal{N}\left(x|\mu,\sigma^2\right). \tag{1.253}$$

Then, by the definition,

$$\mathrm{H}(x) = -\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\ln\mathcal{N}\left(x|\mu,\sigma^2\right)dx. \tag{1.254}$$

The right hand side can be written as

$$-\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx$$

$$=\frac{1}{2}\ln\left(2\pi\sigma^2\right)\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)dx+\frac{1}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2\mathcal{N}\left(x|\mu,\sigma^2\right)dx.$$

$$(1.255)$$

Therefore,

$$\mathrm{H}(x) = \frac{1}{2}\left(1+\ln\left(2\pi\sigma^2\right)\right). \tag{1.256}$$

31

1.36 (Incomplete)

Let f be a strictly convex function. Then, by the definition,

$$f\left(\lambda a + (1-\lambda)b\right) \leq \lambda f(a) + (1-\lambda)f(b), \tag{1.257}$$

where $a \leq b$ and $0 \leq \lambda \leq 1$. Let

$$x = \lambda a + (1-\lambda)b. \tag{1.258}$$

Then, the inequality can be written as

$$f(x) \leq \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \tag{1.259}$$

Let

$$g(x) = \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b) - f(x). \tag{1.260}$$

Then,

$$g(x) \geq 0. \tag{1.261}$$

Additionally, for $x > a$,

$$g(x) = (x-a)\left(\frac{f(b) - f(a)}{b-a} - \frac{f(x) - f(a)}{x-a}\right). \tag{1.262}$$

By the mean value theorem, there exists c and y such that $a \leq c \leq b$, $a \leq y \leq x$ and

$$\begin{aligned}
f'(c) &= \frac{f(b) - f(a)}{b-a}, \\
f'(y) &= \frac{f(x) - f(a)}{x-a}.
\end{aligned} \tag{1.263}$$

Then, for $x > a$, the inequality reduces to

$$f'(y) \leq f'(c). \tag{1.264}$$

1.37

Let $\mathbf{x}$ and $\mathbf{y}$ be two variables. Then, by the definition,

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = -\int\int p(\mathbf{x}, \mathbf{y})\ln p(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y}. \tag{1.265}$$

The right hand side can be written as

$$-\int\int p(\mathbf{x}, \mathbf{y})\left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\right) d\mathbf{x}d\mathbf{y}$$
$$= -\int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x}d\mathbf{y} - \int\left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}\right) \ln p(\mathbf{x}) d\mathbf{x}. \tag{1.266}$$

By the definition, the first term of the right hand side can be written as $H(\mathbf{y}|\mathbf{x})$ and the second term can be written as $H(\mathbf{x})$. Therefore,

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \tag{1.267}$$

1.38

Let f be a strictly convex function. Then, by the definition,

$$f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \tag{1.268}$$

where $0 \leq \lambda \leq 1$. Let us assume that

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i), \tag{1.269}$$

where $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M} \lambda_i = 1. \tag{1.270}$$

Here, let $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M+1} \lambda_i = 1. \tag{1.271}$$

Then, by the definition,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right). \tag{1.272}$$

By the assumption,

$$f\left(\sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right) \leq \sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \tag{1.273}$$

33

Therefore,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \quad (1.274)$$

Thus,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \sum_{i=1}^{M+1} \lambda_i f(x_i). \quad (1.275)$$

Hence, the assumption is proved by induction on M.

1.39

Let x and y be two binary variables where

$$\begin{aligned}
p(x = 0, y = 0) &= \frac{1}{3}, \\
p(x = 0, y = 1) &= \frac{1}{3}, \\
p(x = 1, y = 0) &= 0, \\
p(x = 1, y = 1) &= \frac{1}{3}.
\end{aligned} \quad (1.276)$$

(a)

By the definition,

$$H(x) = -\sum p(x) \ln p(x). \quad (1.277)$$

By the distribution,

$$\begin{aligned}
p(x = 0) &= \frac{2}{3}, \\
p(x = 1) &= \frac{1}{3}.
\end{aligned} \quad (1.278)$$

Therefore,

$$H(x) = \ln 3 - \frac{2}{3} \ln 2. \quad (1.279)$$

34

(b)

By the definition,
$$\mathrm{H}(y) = -\sum p(y) \ln p(y). \tag{1.280}$$

By the distribution,
$$p(y = 0) = \frac{1}{3},$$
$$p(y = 1) = \frac{2}{3}. \tag{1.281}$$

Therefore,
$$\mathrm{H}(y) = \ln 3 - \frac{2}{3} \ln 2. \tag{1.282}$$

(c)

By the definition,
$$\mathrm{H}(y|x) = -\sum p(x, y) \ln p(y|x). \tag{1.283}$$

By the definition,
$$p(y = 0|x = 0) = \frac{p(x = 0, y = 0)}{p(x = 0)},$$
$$p(y = 0|x = 1) = \frac{p(x = 1, y = 0)}{p(x = 1)},$$
$$p(y = 1|x = 0) = \frac{p(x = 0, y = 1)}{p(x = 0)},$$
$$p(y = 1|x = 1) = \frac{p(x = 1, y = 1)}{p(x = 1)}. \tag{1.284}$$

Then, by the distribution,
$$p(y = 0|x = 0) = \frac{1}{2},$$
$$p(y = 0|x = 1) = 0,$$
$$p(y = 1|x = 0) = \frac{1}{2},$$
$$p(y = 1|x = 1) = 1. \tag{1.285}$$

Therefore,
$$\mathrm{H}(y|x) = \frac{2}{3} \ln 2. \tag{1.286}$$

35

(d)

By the definition,
$$H(x|y) = -\sum p(x, y) \ln p(x|y). \tag{1.287}$$

By the definition,
$$\begin{aligned}
p(x = 0|y = 0) &= \frac{p(x = 0, y = 0)}{p(y = 0)}, \\
p(x = 0|y = 1) &= \frac{p(x = 0, y = 1)}{p(y = 1)}, \\
p(x = 1|y = 0) &= \frac{p(x = 1, y = 0)}{p(y = 0)}, \\
p(x = 1|y = 1) &= \frac{p(x = 1, y = 1)}{p(y = 1)}.
\end{aligned} \tag{1.288}$$

Then, by the distribution,
$$\begin{aligned}
p(x = 0|y = 0) &= 1, \\
p(x = 0|y = 1) &= \frac{1}{2}, \\
p(x = 1|y = 0) &= 0, \\
p(x = 1|y = 1) &= \frac{1}{2}.
\end{aligned} \tag{1.289}$$

Therefore,
$$H(x|y) = \frac{2}{3} \ln 2. \tag{1.290}$$

(e)

By the definition,
$$H(x, y) = -\sum p(x, y) \ln p(x, y). \tag{1.291}$$

Therefore,
$$H(x, y) = \ln 3. \tag{1.292}$$

36

(f)

By the definition,

$$\mathrm{I}(x, y) = -\sum p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}. \tag{1.293}$$

By the distribution, the right hand side can be written as

$$\mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y). \tag{1.294}$$

Therefore,

$$\mathrm{I}(x, y) = \ln 3 - \frac{4}{3} \ln 2. \tag{1.295}$$

1.40

Let $\{x_i\}$ be a set of points where $x_i > 0$, and let $\{\lambda_i\}$ be a set of coefficients where $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M} \lambda_i = 1. \tag{1.296}$$

By Jensen's inequality,

$$\sum_{i=1}^{M} \lambda_i \ln x_i \leq \ln \left(\sum_{i=1}^{M} \lambda_i x_i \right). \tag{1.297}$$

Therefore,

$$\prod_{i=1}^{M} x_i^{\lambda_i} \leq \sum_{i=1}^{M} \lambda_i x_i. \tag{1.298}$$

Substituting

$$\lambda_i = \frac{1}{M} \tag{1.299}$$

gives

$$\left(\prod_{i=1}^{M} x_i \right)^{\frac{1}{M}} \leq frac1M \sum_{i=1}^{M} x_i. \tag{1.300}$$

1.41

Let $\mathbf{x}$ and $\mathbf{y}$ be continuous variables. Then, by the definitnion,

$$I(\mathbf{x}, \mathbf{y}) = -\int\int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \tag{1.301}$$

The right hand side can be written as

$$
\begin{aligned}
&-\int\int p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) + \ln \frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \\
&= -\int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y}.
\end{aligned}
\tag{1.302}
$$

By the definition, the first term of the right hand side can be written as $H(\mathbf{x})$ and the second term can be written as $-H(\mathbf{x}|\mathbf{y})$. Therefore,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}). \tag{1.303}$$

By the definition,

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}). \tag{1.304}$$

Thus,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \tag{1.305}$$

2 Probability Distributions

2.1

Let x be a variable such that

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}, \tag{2.1}$$

where $x \in \{0, 1\}$. Then,

$$\sum_x p(x|\mu) = 1. \tag{2.2}$$

By the definition,

$$\begin{aligned} \mathrm{E}\, x &= \mu, \\ \mathrm{E}\, x^2 &= \mu, \end{aligned} \tag{2.3}$$

Since

$$\operatorname{var} x = \mathrm{E}\, x^2 - \left(\mathrm{E}\, x\right)^2, \tag{2.4}$$

we have

$$\operatorname{var} x = \mu(1 - \mu). \tag{2.5}$$

By the definition,

$$\mathrm{H}(x) = -\sum_x p(x|\mu) \ln p(x|\mu). \tag{2.6}$$

Therefore,

$$\mathrm{H}(x) = -\mu \ln \mu - (1 - \mu) \ln(1 - \mu). \tag{2.7}$$

2.2

Let x be a variable such that

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{\frac{1-x}{2}} \left(\frac{1+\mu}{2}\right)^{\frac{1+x}{2}}, \tag{2.8}$$

where $x \in \{-1, 1\}$. Then,

$$\sum_x p(x|\mu) = 1. \tag{2.9}$$

By the definition,

$$\begin{aligned} \mathrm{E}\, x &= \mu, \\ \mathrm{E}\, x^2 &= 1, \end{aligned} \tag{2.10}$$

Since
$$\text{var}\, x = \text{E}\, x^2 - (\text{E}\, x)^2\,,\tag{2.11}$$

we have
$$\text{var}\, x = 1 - \mu^2.\tag{2.12}$$

By the definition,
$$\text{H}(x) = -\sum_x p(x|\mu) \ln p(x|\mu).\tag{2.13}$$

Therefore,
$$\text{H}(x) = -\frac{1-\mu}{2}\ln\frac{1-\mu}{2} - \frac{1+\mu}{2}\ln\frac{1+\mu}{2}.\tag{2.14}$$

2.3

By the definition,
$$\binom{N}{m} = \frac{N!}{m!(N-m)!},$$
$$\binom{N}{m-1} = \frac{N!}{(m-1)!(N-m+1)!}\tag{2.15}$$

Therefore,
$$\binom{N}{m} + \binom{N}{m-1} = \frac{(N-m+1)N! + mN!}{m!(N-m+1)!}.\tag{2.16}$$

By the definition, the right hand side can be written as
$$\frac{(N+1)!}{m!(N+1-m)!} = \binom{N+1}{m}.\tag{2.17}$$

Thus,
$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}.\tag{2.18}$$

Note that
$$1 + x = \sum_{m=0}^{1}\binom{1}{m}x^m.\tag{2.19}$$

Let us assume that
$$(1+x)^N = \sum_{m=0}^{N}\binom{N}{m}x^m.\tag{2.20}$$

Then,

$$(1+x)^{N+1} = \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=0}^{N} \binom{N}{m} x^{m+1}. \tag{2.21}$$

By the result above, the right hand side can be written as

$$\sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=1}^{N+1} \binom{N}{m-1} x^m = 1 + x^{N+1} + \sum_{m=1}^{N} \binom{N+1}{m} x^m. \tag{2.22}$$

Therefore,

$$(1+x)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^m. \tag{2.23}$$

Thus, the assumption is proved by induction on N.

Finally, let m be a variable such that

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}. \tag{2.24}$$

Then

$$\sum_{m=0}^{N} p(m|\mu) = \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m}. \tag{2.25}$$

By the result above, the right hand side can be written as

$$(1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu} \right)^m = (1-\mu)^N \left(1 + \frac{\mu}{1-\mu} \right)^N. \tag{2.26}$$

Therefore,

$$\sum_{m=0}^{N} p(m|\mu) = 1. \tag{2.27}$$

2.4

Let m be a variable such that

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}. \tag{2.28}$$

Then

$$\mathrm{E}\,m = \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m}. \tag{2.29}$$

Taking the derivative of

$$\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1 \tag{2.30}$$

with respect to μ gives

$$\sum_{m=0}^{N} m \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m} - \sum_{m=0}^{N}(N-m)\binom{N}{m}\mu^m(1-\mu)^{N-m-1} = 0. \tag{2.31}$$

The first term of the left hand side can be written as $\frac{1}{\mu}\,\mathrm{E}\,m$. Since

$$(N-m)\binom{N}{m} = N\binom{N-1}{m}, \tag{2.32}$$

the second term of the left hand side can be written as

$$-N\sum_{m=0}^{N-1}\binom{N-1}{m}\mu^m(1-\mu)^{N-m-1} = -N. \tag{2.33}$$

Therefore,

$$\mathrm{E}\,m = N\mu. \tag{2.34}$$

Taking the second derivative of

$$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 1 \tag{2.35}$$

with respect to μ gives

$$\begin{aligned}
&\sum_{m=0}^{N} m(m-1)\binom{N}{m}\mu^{m-2}(1-\mu)^{N-m}\\
&-2\sum_{m=0}^{N} m(N-m)\binom{N}{m}\mu^{m-1}(1-\mu)^{N-m-1}\\
&+\sum_{m=0}^{N}(N-m)(N-m-1)\binom{N}{m}\mu^m(1-\mu)^{N-m-2} = 0.
\end{aligned} \tag{2.36}$$

The first term of the left hand side can be written as $\frac{1}{\mu^2} \mathrm{E}\, m(m-1)$. Since

$$m(N-m)\binom{N}{m} = N(N-1)\binom{N-2}{m-1},$$

$$(N-m)(N-m-1)\binom{N}{m} = N(N-1)\binom{N-2}{m},$$

(2.37)

the second and third term of the left hand side can be written as

$$-2N(N-1)\sum_{m=1}^{N-1}\binom{N-2}{m-1}\mu^{m-1}(1-\mu)^{N-m-1} = -2N(N-1),$$

$$N(N-1)\sum_{m=0}^{N}\binom{N-2}{m}\mu^{m}(1-\mu)^{N-m-2} = N(N-1).$$

(2.38)

Therefore,

$$\mathrm{E}\, m(m-1) = N(N-1)\mu^2.$$

(2.39)

Thus, since

$$\mathrm{var}\, m = \mathrm{E}\, m(m-1) + \mathrm{E}\, m - (\mathrm{E}\, m)^2,$$

(2.40)

we have

$$\mathrm{var}\, m = N\mu(1-\mu).$$

(2.41)

2.5

By the definition,

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}\exp(-x)dx \int_0^\infty y^{b-1}\exp(-y)dy.$$

(2.42)

By the transformation $t = x + y$, the right hand side can be written as

$$\int_0^\infty x^{a-1}\left(\int_x^\infty (t-x)^{b-1}\exp(-t)dt\right)dx$$

$$= \int_0^\infty \left(\int_0^t x^{a-1}(t-x)^{b-1}dx\right)\exp(-t)dt.$$

(2.43)

By the transformation $x = t\mu$, the right hand side can be written as

$$\int_0^\infty \left(\int_0^1 (t\mu)^{a-1}t^{b-1}(1-\mu)^{b-1}td\mu\right)\exp(-t)dt$$

$$= \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu \int_0^\infty t^{a+b-1}\exp(-t)dt.$$

(2.44)

By the definition, the second integral of the right hand side can be written as $\Gamma(a + b)$. Therefore,

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.45}$$

2.6

Let μ be a variable such that

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}. \tag{2.46}$$

Then

$$\mathrm{E}\,\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1}d\mu,$$

$$\mathrm{E}\,\mu^2 = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1-\mu)^{b-1}d\mu. \tag{2.47}$$

Since

$$\int_0^1 \mu^a(1-\mu)^{b-1}d\mu = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)},$$

$$\int_0^1 \mu^{a+1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}, \tag{2.48}$$

we have

$$\mathrm{E}\,\mu = \frac{a}{a+b},$$

$$\mathrm{E}\,\mu^2 = \frac{a(a+1)}{(a+b)(a+b+1)}. \tag{2.49}$$

Since

$$\mathrm{var}\,\mu = \mathrm{E}\,\mu^2 - (\mathrm{E}\mu)^2, \tag{2.50}$$

we have

$$\mathrm{var}\,\mu = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.51}$$

Since

$$\frac{\partial}{\partial\mu}p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\left(\frac{a-1}{\mu} - \frac{b-1}{1-\mu}\right), \tag{2.52}$$

we have

$$\mathrm{mode}\,\mu = \frac{a-1}{a+b-2}. \tag{2.53}$$

2.7

Let m and l be a variable such that

$$p(m, l | \mu) = \binom{m + l}{m} \mu^m (1 - \mu)^l, \tag{2.54}$$

where

$$p(\mu | a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1}. \tag{2.55}$$

By 2.6,

$$\mathrm{E}(\mu | a, b) = \frac{a}{a + b}. \tag{2.56}$$

Note that

$$\mu_{\mathrm{ML}} = \frac{m}{m + l}. \tag{2.57}$$

Since

$$p(\mu | m, l, a, b) \propto p(m, l | \mu) p(\mu | a, b), \tag{2.58}$$

we have

$$p(\mu | m, l, a, b) = \frac{\Gamma(m + l + a + b)}{\Gamma(m + a)\Gamma(l + b)} \mu^{m+a-1}(1 - \mu)^{l+b-1}. \tag{2.59}$$

Therefore, by 2.6,

$$\mathrm{E}(\mu | m, l, a, b) = \frac{m + a}{m + l + a + b}. \tag{2.60}$$

Thus,

$$\mathrm{E}(\mu | m, l, a, b) = \lambda \mu_{\mathrm{ML}} + (1 - \lambda) \mathrm{E}(\mu | a, b), \tag{2.61}$$

where

$$\lambda = \frac{m + l}{m + l + a + b}. \tag{2.62}$$

2.8

Let x and y be variables. Then, by the definition,

$$\mathrm{E}\, x = \int x p(x) dx. \tag{2.63}$$

The right hand side can be written as

$$\int x \left(\int p(x, y) dy \right) dx = \int \left(\int x p(x | y) dx \right) p(y) dy. \tag{2.64}$$

Therefore,
$$\mathrm{E}\,x = \mathrm{E}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.65}$$

Additionally, by the definition,
$$\operatorname{var} x = \mathrm{E}\left(x - \mathrm{E}\,x\right)^2. \tag{2.66}$$

By the result above, the right hand side can be written as
$$
\begin{aligned}
&\mathrm{E}_y\left(\mathrm{E}_x\left((x - \mathrm{E}_x(x|y) + \mathrm{E}_x(x|y) - \mathrm{E}\,x)^2\,|y\right)\right) \\
={}&\mathrm{E}_y\left(\mathrm{E}_x\left((x - \mathrm{E}_x(x|y))^2\,|y\right)\right) \\
&+ 2\,\mathrm{E}_y\left(((\mathrm{E}_x(x|y) - \mathrm{E}\,x)\,\mathrm{E}_x\left(x - \mathrm{E}_x(x|y)\right)|y)\right) \\
&+ \mathrm{E}_y\left((\mathrm{E}_x(x|y) - \mathrm{E}\,x)^2\,|y\right).
\end{aligned}
\tag{2.67}
$$

Let us look at each term of the right hand side. By the definition, the first term can be written as $\mathrm{E}_y\left(\operatorname{var}_x(x|y)\right)$. The second term can be written as
$$2\,\mathrm{E}_y\left((\mathrm{E}_x(x|y) - \mathrm{E}\,x)\left(\mathrm{E}_x(x|y) - \mathrm{E}_x(x|y)\right)\right) = 0. \tag{2.68}$$

By the result above, the third term can be written as
$$\mathrm{E}_y\left(\mathrm{E}_x(x|y) - \mathrm{E}_y\left(\mathrm{E}_x(x|y)\right)\right)^2 = \operatorname{var}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.69}$$

Therefore,
$$\operatorname{var} x = \mathrm{E}_y\left(\operatorname{var}_x(x|y)\right) + \operatorname{var}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.70}$$

2.9

For a vector $\boldsymbol{\mu}$ in 2 dimensions, 2.5 gives
$$\int_{\substack{\mu_1+\mu_2=1 \\ \mu_1\geq 0,\mu_2\geq 0}} \mu_1^{\alpha_1-1}\mu_2^{\alpha_2-1}d\boldsymbol{\mu} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

For a vector $\boldsymbol{\mu}$ in M dimensions, let us assume that
$$\int_{\substack{\sum_{k=1}^{M}\mu_k=1 \\ \mu_k\geq 0}} \prod_{k=1}^{M} \mu_k^{\alpha_k-1}d\boldsymbol{\mu} = \frac{\prod_{k=1}^{M}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{M}\alpha_k)}.$$

Under the constraint
$$\sum_{k=1}^{M+1} \mu_k = 1, \tag{2.71}$$

we have

$$\int_0^{1-\sum_{k=1}^{M-1}\mu_k} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\mu_{M+1}$$

$$= \left(\prod_{k=1}^{M-1} \mu_k^{\alpha_k-1}\right) \int_0^{1-\sum_{k=1}^{M-1}\mu_k} \mu_{M+1}^{\alpha_{M+1}-1} \left(1 - \sum_{k=1}^{M-1} \mu_k - \mu_{M+1}\right)^{\alpha_M-1} d\mu_{M+1}. \tag{2.72}$$

By the transformation

$$\mu'_{M+1} = \frac{\mu_{M+1}}{1 - \sum_{k=1}^{M-1} \mu_k}, \tag{2.73}$$

the integral of the right hand side can be written as

$$\int_0^1 \left(\left(1 - \sum_{k=1}^{M-1} \mu_k\right) \mu'_{M+1}\right)^{\alpha_{M+1}-1} \left(\left(1 - \sum_{k=1}^{M-1} \mu_k\right)(1 - \mu'_{M+1})\right)^{\alpha_M-1} \left(1 - \sum_{k=1}^{M-1} \mu_k\right) d\mu'_{M+1}$$

$$= \left(1 - \sum_{k=1}^{M-1} \mu_k\right)^{\alpha_M+\alpha_{M+1}-1} \int_0^1 \mu'_{M+1}{}^{\alpha_{M+1}-1}(1 - \mu'_{M+1})^{\alpha_M-1} d\mu'_{M+1}. \tag{2.74}$$

By 2.5, the integral of the right hand side can be written as

$$\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \tag{2.75}$$

Therefore,

$$\int_0^{1-\sum_{k=1}^{M-1}\mu_k} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\mu_{M+1} = \left(\prod_{k=1}^{M-1} \mu_k^{\alpha_k-1}\right) \left(1 - \sum_{k=1}^{M-1} \mu_k\right)^{\alpha_M+\alpha_{M+1}-1} \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \tag{2.76}$$

By the assumption,

$$\int_{\substack{\sum_{k=1}^{M}\mu_k=1 \\ \mu_k\geq 0}} \left(\prod_{k=1}^{M-1} \mu_k^{\alpha_k-1}\right) \mu_M^{\alpha_M+\alpha_{M+1}-1} d\boldsymbol{\mu} = \frac{(\prod_{k=1}^{M-1}\Gamma(\alpha_k))\Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{k=1}^{M+1}\alpha_k)}.$$

Thus, for a vector $\boldsymbol{\mu}$ in $M + 1$ dimensions,

$$\int_{\substack{\sum_{k=1}^{M+1}\mu_k=1 \\ \mu_k\geq 0}} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\boldsymbol{\mu} = \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})} \frac{(\prod_{k=1}^{M-1}\Gamma(\alpha_k))\Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{k=1}^{M+1}\alpha_k)}.$$

47

The right hand side can be written as

$$\frac{\prod_{k=1}^{M+1} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{M+1} \alpha_k)}. \tag{2.77}$$

Hence, the assumption is proved by induction on M.

2.10

Let $\boldsymbol{\mu}$ be a variable such that

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}. \tag{2.78}$$

Then

$$\mathrm{E}\,\mu_j = \int \mu_j p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu},$$

$$\mathrm{E}\,\mu_j^2 = \int \mu_j^2 p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \tag{2.79}$$

$$\mathrm{E}\,\mu_j \mu_l = \int \mu_j \mu_l p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}.$$

If $j \neq l$, then the right hand sides can be written as

$$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j + 1)}{\Gamma(\alpha_j)} \prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + 1)} = \frac{\alpha_j}{\sum_{k=1}^{K} \alpha_k},$$

$$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j + 2)}{\Gamma(\alpha_j)} \prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + 2\right)} = \frac{\alpha_j(\alpha_j + 1)}{\sum_{k=1}^{K} \alpha_k(\sum_{k=1}^{K} \alpha_k + 1)}, \tag{2.80}$$

$$\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j + 1)\Gamma(\alpha_l + 1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)} \prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + 2\right)} = \frac{\alpha_j \alpha_l}{\sum_{k=1}^{K} \alpha_k(\sum_{k=1}^{K} \alpha_k + 1)}.$$

Therefore,

$$\mathrm{E}\,\mu_j = \frac{\alpha_j}{\sum_{k=1}^{K} \alpha_k}.$$

$$\mathrm{E}\,\mu_j^2 = \frac{\alpha_j(\alpha_j + 1)}{\sum_{k=1}^{K} \alpha_k \left(\sum_{k=1}^{K} \alpha_k + 1\right)}, \tag{2.81}$$

$$\mathrm{E}\,\mu_j \mu_l = \frac{\alpha_j \alpha_l}{\sum_{k=1}^{K} \alpha_k \left(\sum_{k=1}^{K} \alpha_k + 1\right)}.$$

Since
$$\operatorname{var} \mu_j = \operatorname{E} \mu_j^2 - \left(\operatorname{E} \mu_j\right)^2,$$
$$\operatorname{cov}\left(\mu_j, \mu_l\right) = \operatorname{E} \mu_j \mu_l - \operatorname{E} \mu_j \operatorname{E} \mu_l, \tag{2.82}$$

we have
$$\operatorname{var} \mu_j = \frac{\alpha_j\left(\sum_{k=1}^K \alpha_k - \alpha_j\right)}{\sum_{k=1}^K \alpha_k)^2\left(\sum_{k=1}^K \alpha_k + 1\right)},$$
$$\operatorname{cov}\left(\mu_j, \mu_l\right) = -\frac{\alpha_j \alpha_l}{\left(\sum_{k=1}^K \alpha_k\right)^2\left(\sum_{k=1}^K \alpha_k + 1\right)}. \tag{2.83}$$

2.11

Let $\boldsymbol{\mu}$ be a variable such that

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}. \tag{2.84}$$

Then
$$\operatorname{E} \ln \mu_j = \int \left(\ln \mu_j\right) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \tag{2.85}$$

Since

$$\frac{\partial}{\partial \alpha_j} p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \left(\frac{\Gamma'\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} - \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} + \ln \mu_j\right) p(\boldsymbol{\mu}|\boldsymbol{\alpha}), \tag{2.86}$$

we have

$$\operatorname{E} \ln \mu_j = \frac{\partial}{\partial \alpha_j} \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} + \left(\psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k\right)\right) \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \tag{2.87}$$

where
$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \tag{2.88}$$

Therefore,
$$\operatorname{E} \ln \mu_j = \psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k\right). \tag{2.89}$$

2.12

Let x be a variable such that

$$p(x|a, b) = \frac{1}{b - a},\tag{2.90}$$

where $a < b$. Then

$$\int_a^b p(x|a, b)dx = 1.\tag{2.91}$$

Note that

$$\mathrm{E}\, x = \int_a^b xp(x|a, b)dx,$$
$$\mathrm{E}\, x^2 = \int_a^b x^2 p(x|a, b)dx.\tag{2.92}$$

The right hand sides can be written as

$$\frac{1}{b - a}\int_a^b xdx = \frac{1}{2}(a + b),$$
$$\frac{1}{b - a}\int_a^b x^2 dx = \frac{1}{3}\left(a^2 + ab + b^2\right).\tag{2.93}$$

Therefore,

$$\mathrm{E}\, x = \frac{1}{2}(a + b),$$
$$\mathrm{E}\, x^2 = \frac{1}{3}\left(a^2 + ab + b^2\right).\tag{2.94}$$

Since

$$\mathrm{var}\, x = \mathrm{E}\, x^2 - (\mathrm{E}\, x)^2,\tag{2.95}$$

we have

$$\mathrm{var}\, x = \frac{1}{12}(b - a)^2.\tag{2.96}$$

2.13

Let $\mathbf{x}$ be a variable in D dimensions and

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$
$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}).\tag{2.97}$$

Then, by the definition,

$$\text{KL}(p||q) = -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \tag{2.98}$$

Since

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} \left(|\det \mathbf{L}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m})\right)}{(2\pi)^{-\frac{D}{2}} \left(|\det \boldsymbol{\Sigma}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}, \tag{2.99}$$

The right hand side can be written as

$$\frac{1}{2} \ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$$

$$+ \frac{1}{2} \int (\mathbf{x} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \tag{2.100}$$

$$- \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.$$

Let us look at each term. Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1, \tag{2.101}$$

the first term can be written as $\frac{1}{2} \ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right|$. Since

$$(\mathbf{x} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) = (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m}), \tag{2.102}$$

the second term can be written as

$$\frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$$

$$+ (\boldsymbol{\mu} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1} \int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \tag{2.103}$$

$$+ \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m})^{\mathsf{T}} \mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.$$

Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1,$$

$$\int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\mu}, \tag{2.104}$$

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\Sigma},$$

it can be written as

$$\frac{1}{2}\operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right) + \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^{\mathsf{T}}\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}). \tag{2.105}$$

Since

$$\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.106}$$

the third term can be written as

$$-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = -\frac{D}{2}. \tag{2.107}$$

Therefore,

$$\mathrm{KL}(p||q) = \frac{1}{2}\left(\ln\left|\frac{\det\mathbf{L}}{\det\boldsymbol{\Sigma}}\right| + \operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right) + (\boldsymbol{\mu}-\mathbf{m})^{\mathsf{T}}\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) - D\right). \tag{2.108}$$

2.14

Let $\mathbf{x}$ be a variable in D dimensions and

$$L(p(\mathbf{x})) = -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x} + \lambda\left(\int p(\mathbf{x})d\mathbf{x} - 1\right)$$
$$+ \mathbf{l}^{\mathsf{T}}\left(\int \mathbf{x}p(\mathbf{x})d\mathbf{x} - \boldsymbol{\mu}\right) + \mathbf{m}^{\mathsf{T}}\left(\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}p(\mathbf{x})d\mathbf{x} - \boldsymbol{\Sigma}\right)\mathbf{m}. \tag{2.109}$$

Then

$$\frac{\delta L(p(\mathbf{x}))}{\delta p(\mathbf{x})} = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^{\mathsf{T}}\mathbf{x} + \mathbf{m}^{\mathsf{T}}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{m}. \tag{2.110}$$

Setting the left hand side to zero gives

$$p(\mathbf{x}) = \exp\left(-1 + \lambda + \mathbf{l}^{\mathsf{T}}\mathbf{x} + \mathbf{m}^{\mathsf{T}}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{m}\right), \tag{2.111}$$

so that

$$p(\mathbf{x}) = \exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{M}\mathbf{l} + (\mathbf{x}-\boldsymbol{\mu}-\mathbf{M}\mathbf{l})^{\mathsf{T}}\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{M}\mathbf{l})\right), \tag{2.112}$$

where

$$\mathbf{M} = \left(\mathbf{m}\mathbf{m}^{\mathsf{T}}\right)^{-1}. \tag{2.113}$$

Substituting it to

$$\int p(\mathbf{x})d\mathbf{x} = 1,$$

$$\int \mathbf{x}p(\mathbf{x})d\mathbf{x} = \boldsymbol{\mu}, \qquad (2.114)$$

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}p(\mathbf{x})d\mathbf{x} = \boldsymbol{\Sigma},$$

and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml} \qquad (2.115)$$

gives

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)\int \exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{y}\right)d\mathbf{y} = 1,$$

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)\int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{Ml})\exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{y}\right)d\mathbf{y} = \boldsymbol{\mu},$$

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)\int (\mathbf{y} + \mathbf{Ml})(\mathbf{y} + \mathbf{Ml})^{\mathsf{T}}\exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{y}\right)d\mathbf{y} = \boldsymbol{\Sigma}.$$

$$(2.116)$$

Since

$$\int \exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{y}\right)d\mathbf{y} = \left(\Gamma\left(\frac{1}{2}\right)\right)^{D},$$

$$\int \mathbf{y}\exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{y}\right)d\mathbf{y} = \mathbf{0}, \qquad (2.117)$$

$$\int \mathbf{y}\mathbf{y}^{\mathsf{T}}\exp\left(-\mathbf{y}^{\mathsf{T}}\mathbf{y}\right)d\mathbf{y} = \Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1}\mathbf{I},$$

they can be written as

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}(\det\mathbf{M})^{\frac{1}{2}} = 1,$$

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)(\boldsymbol{\mu} + \mathbf{Ml})\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}(\det\mathbf{M})^{\frac{1}{2}} = \boldsymbol{\mu},$$

$$\exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right)\left(\Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1}\mathbf{M} + \mathbf{Ml}(\mathbf{Ml})^{\mathsf{T}}\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}\right)(\det\mathbf{M})^{\frac{1}{2}} = \boldsymbol{\Sigma}.$$

$$(2.118)$$

Therefore,

$$\lambda = 1 - \frac{D}{2} \ln \pi - \frac{1}{2} \ln(\det \mathbf{M}),$$
$$\mathbf{l} = \mathbf{0},$$
$$\mathbf{M} = 2\mathbf{\Sigma}. \tag{2.119}$$

Thus,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} (\det \mathbf{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{2.120}$$

2.15

Let $\mathbf{x}$ be a variable in D dimensions such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}). \tag{2.121}$$

Then, by the definition,

$$\mathrm{H}(\mathbf{x}) = -\int \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) d\mathbf{x}. \tag{2.122}$$

The right hand side can be written as

$$-\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln|\det \mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) d\mathbf{x}$$
$$= \left(\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln|\det \mathbf{\Sigma}|\right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x}$$
$$+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x}. \tag{2.123}$$

Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} = 1,$$
$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} = \mathbf{\Sigma}, \tag{2.124}$$

the first and second term of the right hand side can be written as

$$\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln|\det \mathbf{\Sigma}| \tag{2.125}$$

and

$$\frac{1}{2} \operatorname{tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \right) = \frac{D}{2}. \tag{2.126}$$

Therefore,

$$\mathrm{H}(\mathbf{x}) = \frac{D}{2} \left(1 + \ln(2\pi) \right) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}|. \tag{2.127}$$

2.16

Let x be a variable such that

$$x = x_1 + x_2, \tag{2.128}$$

where

$$\begin{aligned} p(x_1) &= \mathcal{N} \left(x_1 | \mu_1, \tau_1^{-1} \right), \\ p(x_2) &= \mathcal{N} \left(x_2 | \mu_2, \tau_2^{-1} \right). \end{aligned} \tag{2.129}$$

By marginalisation,

$$p(x) = \int_{-\infty}^{\infty} p(x|x_2) p(x_2) dx_2. \tag{2.130}$$

The right hand side can be written as

$$\int_{-\infty}^{\infty} \mathcal{N} \left(x | \mu_1 + x_2, \tau_1^{-1} \right) \mathcal{N} \left(x_2 | \mu_2, \tau_2^{-1} \right) dx_2$$
$$= \int_{-\infty}^{\infty} \left(\frac{\tau_1}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\tau_1}{2} (x - \mu_1 - x_2)^2 \right) \left(\frac{\tau_2}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\tau_2}{2} (x_2 - \mu_2)^2 \right) dx_2. \tag{2.131}$$

The logarithm of the integrand except the terms independent of x and z is given by

$$-\frac{\tau_1 + \tau_2}{2} \left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2$$
$$+ \frac{\tau_1 + \tau_2}{2} \left(\frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2$$
$$= -\frac{\tau_1 + \tau_2}{2} \left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2. \tag{2.132}$$

Therefore,
$$p(x) = \mathcal{N}\left(x|\mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}\right). \tag{2.133}$$

Thus, by 1.35,

$$H(x) = \frac{1}{2}\left(1 + \ln(2\pi) + \ln\left(\tau_1^{-1} + \tau_2^{-1}\right)\right). \tag{2.134}$$

2.17

Let $\boldsymbol{\Sigma}$ be a matrix and

$$\begin{aligned}
\mathbf{S} &= \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right), \\
\mathbf{A} &= \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} - \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right).
\end{aligned} \tag{2.135}$$

Then
$$\boldsymbol{\Sigma}^{-1} = \mathbf{S} + \mathbf{A}. \tag{2.136}$$

Therefore,

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.137}$$

The second term of the right hand side can be written as

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.138}$$

The second term of the right hand side can be written as

$$-\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.139}$$

Thus,
$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = 0. \tag{2.140}$$

Hence
$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.141}$$

2.18

Let $\mathbf{\Sigma}$ be a $D \times D$ real symmetric matrix such that

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \tag{2.142}$$

where $i = 1, \cdots, D$ and $\mathbf{u}_i$ are unit vectors. Taking the inner product with $\overline{\mathbf{u}_i}$ on both sides gives

$$\overline{\mathbf{u}_i}^\mathsf{T}\mathbf{\Sigma}\mathbf{u}_i = \lambda_i. \tag{2.143}$$

Since $\mathbf{\Sigma}$ is real and symmetric, the left hand side can be written as

$$\overline{\mathbf{u}_i}^\mathsf{T}\overline{\mathbf{\Sigma}}^\mathsf{T}\mathbf{u}_i = \left(\overline{\mathbf{\Sigma}}\overline{\mathbf{u}_i}\right)^\mathsf{T}\mathbf{u}_i. \tag{2.144}$$

The right hand side can be writtet as

$$\overline{\lambda}_i\overline{\mathbf{u}}_i^\mathsf{T}\mathbf{u}_i = \overline{\lambda}_i. \tag{2.145}$$

Therefore,

$$\lambda_i = \overline{\lambda}_i. \tag{2.146}$$

Additionally, for $i \neq j$, taking the inner product with $\mathbf{u}_j$ on n both sides of the original equation gives

$$\mathbf{u}_j^\mathsf{T}\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_j^\mathsf{T}\mathbf{u}_i. \tag{2.147}$$

Since $\mathbf{\Sigma}$ is symmetric, the left hand side can be written as

$$\mathbf{u}_j^\mathsf{T}\mathbf{\Sigma}^\mathsf{T}\mathbf{u}_i = \left(\mathbf{\Sigma}\mathbf{u}_j\right)^\mathsf{T}\mathbf{u}_i. \tag{2.148}$$

The right hand side can be written as $\lambda_j\mathbf{u}_j^\mathsf{T}\mathbf{u}_i$. Therefore,

$$\lambda_i\mathbf{u}_j^\mathsf{T}\mathbf{u}_i = \lambda_j\mathbf{u}_j^\mathsf{T}\mathbf{u}_i. \tag{2.149}$$

Thus, if $\lambda_i \neq \lambda_j$, then

$$\mathbf{u}_j^\mathsf{T}\mathbf{u}_i = 0. \tag{2.150}$$

2.19

Let $\mathbf{\Sigma}$ be a $D \times D$ real symmetric matrix such that

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \tag{2.151}$$

57

where $i = 1, \cdots, D$ and $\mathbf{u}_i$ are unit vectors. Let

$$\begin{aligned}\boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_1, \cdots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].\end{aligned} \tag{2.152}$$

Then

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}. \tag{2.153}$$

By 2.18,

$$\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}. \tag{2.154}$$

Therefore,

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}, \\ \boldsymbol{\Sigma}^{-1} &= \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathsf{T}},\end{aligned} \tag{2.155}$$

Thus,

$$\begin{aligned}\boldsymbol{\Sigma} &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}, \\ \boldsymbol{\Sigma}^{-1} &= \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}.\end{aligned} \tag{2.156}$$

2.20

Let $\boldsymbol{\Sigma}$ be a $D \times D$ real symmetric matrix such that

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.157}$$

where $i = 1, \cdots, D$ and $\mathbf{u}_i$ are unit vectors. Let

$$\begin{aligned}\boldsymbol{\Lambda} &= \operatorname{diag}(\lambda_1, \cdots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].\end{aligned} \tag{2.158}$$

By 2.19,

$$\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{a} = \mathbf{b}^{\mathsf{T}}\boldsymbol{\Lambda}\mathbf{b}, \tag{2.159}$$

where

$$\mathbf{b} = \mathbf{U}^{\mathsf{T}}\mathbf{a}. \tag{2.160}$$

The right hand side can be written as $\sum_{i=1}^{D} \lambda_i b_i^2$. Therefore, the necessary and sufficient condition for

$$\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{a} > 0 \tag{2.161}$$

for any real vector $\mathbf{a}$ is

$$\lambda_i > 0. \tag{2.162}$$

2.21

Let $\mathbf{\Sigma}$ be a $D \times D$ real symmetric matrix. Then the number of independent parameters is $\frac{D(D+1)}{2}$.

2.22

Let $\mathbf{\Sigma}$ be a $D \times D$ symmetric matrix and

$$\mathbf{\Sigma}\mathbf{\Lambda} = \mathbf{I}. \tag{2.163}$$

Taking the transpose of the both sides gives

$$\mathbf{\Lambda}^\mathsf{T}\mathbf{\Sigma} = \mathbf{I}. \tag{2.164}$$

Therefore,

$$\mathbf{\Lambda}^\mathsf{T} = \mathbf{\Lambda}. \tag{2.165}$$

2.23

Let $\mathbf{\Sigma}$ be a $D \times D$ real symmetric matrix such that

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.166}$$

where $i = 1, \cdots, D$ and $\mathbf{u}_i$ are unit vectors. Let

$$\begin{aligned}
\mathbf{\Lambda}' &= \mathrm{diag}\left(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_D^{-\frac{1}{2}}\right), \\
\mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].
\end{aligned} \tag{2.167}$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{U}\mathbf{\Lambda}'\mathbf{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \tag{2.168}$$

By the transformation

$$\mathbf{y} = \mathbf{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu}) \tag{2.169}$$

and the property

$$\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}, \tag{2.170}$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left|\det\left(\mathbf{U}\boldsymbol{\Lambda'}^{-1}\right)\right| d\mathbf{y} = |\det\boldsymbol{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \qquad (2.171)$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det\boldsymbol{\Sigma}|^{\frac{1}{2}}\Delta^D V_D, \qquad (2.172)$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \qquad (2.173)$$

2.24

Let

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

be a partitioned matrix where $\mathbf{A}$ is a square matrix and $\mathbf{D}$ is an invertible matrix. By an LDU decomposition, we have

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}.$$

Therefore,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}.$$

Thus,

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

2.25

Let $\mathbf{x}$ be a variable in D dimensions such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad (2.174)$$

where
$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \\ \mathbf{x}_c \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bc} \\ \boldsymbol{\Sigma}_{ca} & \boldsymbol{\Sigma}_{cb} & \boldsymbol{\Sigma}_{cc} \end{bmatrix}.$$

Let
$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{2.175}$$

where
$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} & \boldsymbol{\Lambda}_{ac} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} & \boldsymbol{\Lambda}_{bc} \\ \boldsymbol{\Lambda}_{ca} & \boldsymbol{\Lambda}_{cb} & \boldsymbol{\Lambda}_{cc} \end{bmatrix}.$$

Then

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\
=&-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathsf{T}} \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathsf{T}} \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathsf{T}} \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\
&-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^{\mathsf{T}} \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^{\mathsf{T}} \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^{\mathsf{T}} \boldsymbol{\Lambda}_{bc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\
&-\frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^{\mathsf{T}} \boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^{\mathsf{T}} \boldsymbol{\Lambda}_{cb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^{\mathsf{T}} \boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c).
\end{aligned}
\tag{2.176}
$$

Excluding the terms independent of $\mathbf{x}_a$, the right hand side can be written as

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^{\mathsf{T}} \boldsymbol{\Sigma}_{a|b,c}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \tag{2.177}$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} (\mathbf{x}_c - \boldsymbol{\mu}_c), \\
\boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}.
\end{aligned}
\tag{2.178}
$$

Therefore,
$$p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \tag{2.179}$$

Multiplying both sides by $p(\mathbf{x}_c)$ and integrating both sides with respect to $\mathbf{x}_c$ gives

$$p(\mathbf{x}_a|\mathbf{x}_b) = \int \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}) p(\mathbf{x}_c) d\mathbf{x}_c. \tag{2.180}$$

Thus,
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \tag{2.181}$$

61

where
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\mu}_c,$$
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}. \tag{2.182}$$

2.26

We have
$$\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})$$
$$=\mathbf{I} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D} + \mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D} \tag{2.183}$$
$$- \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D}.$$

The right hand side except the first term can be written as
$$\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{I} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\right)\right)\mathbf{D}$$
$$=\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)\mathbf{C}\right)\mathbf{D}. \tag{2.184}$$

The right hand sidde can be written as
$$\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \mathbf{C}\right)\mathbf{D} = \mathbf{O}. \tag{2.185}$$

Therefore,
$$\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) = \mathbf{I}. \tag{2.186}$$

Thus,
$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}. \tag{2.187}$$

2.27

Let $\mathbf{x}$ and $\mathbf{z}$ be two variables. Then
$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \int\int(\mathbf{x} + \mathbf{z})p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.188}$$

The right hand side can be written as
$$\int\mathbf{x}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{z}\right)d\mathbf{x} + \int\mathbf{z}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{x}\right)d\mathbf{z} = \int\mathbf{x}p(\mathbf{x})d\mathbf{x} + \int\mathbf{z}p(\mathbf{z})d\mathbf{z}. \tag{2.189}$$

The right hand side can be written as $\mathrm{E}\,\mathbf{x} + \mathrm{E}\,\mathbf{z}$. Therefore,

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \mathrm{E}\,\mathbf{x} + \mathrm{E}\,\mathbf{z}. \tag{2.190}$$

Additionally,

$$\mathrm{cov}(\mathbf{x} + \mathbf{z}) = \int\int (\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z}))(\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z}))^\mathsf{T} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.191}$$

The right hand side can be written as

$$\int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^\mathsf{T} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^\mathsf{T} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}$$
$$+ \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^\mathsf{T} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^\mathsf{T} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.192}$$

The first and fourth terms can be written as $\mathrm{cov}\,\mathbf{x}$ and $\mathrm{cov}\,\mathbf{z}$. If $\mathbf{x}$ and $\mathbf{z}$ are independent, the second and third terms can be written as

$$\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})\,p(\mathbf{x})d\mathbf{x} \int (\mathbf{z} - \mathrm{E}\,\mathbf{z})^\mathsf{T} p(\mathbf{z})d\mathbf{z} = \mathbf{O},$$
$$\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})\,p(\mathbf{z})d\mathbf{z} \int (\mathbf{x} - \mathrm{E}\,\mathbf{x})^\mathsf{T} p(\mathbf{x})d\mathbf{x} = \mathbf{O}. \tag{2.193}$$

Therefore,

$$\mathrm{cov}(\mathbf{x} + \mathbf{z}) = \mathrm{cov}\,\mathbf{x} + \mathrm{cov}\,\mathbf{z}. \tag{2.194}$$

2.28

Let $\mathbf{x}$ and $\mathbf{y}$ be Gaussian variables and

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

where

$$\mathrm{E}\,\mathbf{z} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}$$

and

$$\mathrm{cov}\,\mathbf{z} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix}.$$

Then, by 2.29,

$$(\text{cov } \mathbf{z})^{-1} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA} & -\mathbf{A}^\mathsf{T}\mathbf{L} \\ -\mathbf{LA} & \mathbf{L} \end{bmatrix}.$$

Then, $\ln p(\mathbf{z})$ except the terms independent of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA}\right)(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})$$

$$+ \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{LA}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})$$

$$= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA}\right)(\mathbf{x} - \boldsymbol{\mu}) +$$

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right)^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right)$$

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{LA}(\mathbf{x} - \boldsymbol{\mu}).$$

$$(2.195)$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right)^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right). \qquad (2.196)$$

Therefore,

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}\right). \qquad (2.197)$$

2.29

Let

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA} & -\mathbf{A}^\mathsf{T}\mathbf{L} \\ -\mathbf{LA} & \mathbf{L} \end{bmatrix}.$$

Then, by 2.24,

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix}.$$

2.30

Let

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix}.$$

Then

$$\mathbf{R}^{-1} \begin{bmatrix} \mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\mathsf{T}\mathbf{Lb} \\ \mathbf{Lb} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}.$$

2.31

Let $\mathbf{y}$ be a variable such that

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{2.198}$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}\right), \\ p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z}\right). \end{aligned} \tag{2.199}$$

By marginalisation,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{2.200}$$

The right hand side can be written as

$$\int \mathcal{N}\left(\mathbf{y}|\mathbf{x} + \boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z}\right) \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x}\right) d\mathbf{x}. \tag{2.201}$$

The logarithm of the integrand except the terms independent of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right)^\mathsf{T} \boldsymbol{\Sigma}_\mathbf{z}^{-1}\left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}). \tag{2.202}$$

The first and second order terms can be written as

$$-\mathbf{x}^\mathsf{T}\left(\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} - \boldsymbol{\Sigma}_\mathbf{x}^{-1}\boldsymbol{\mu}_\mathbf{x}\right) + \mathbf{y}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} = \mathbf{u}^\mathsf{T}\mathbf{v} \tag{2.203}$$

and

$$-\frac{1}{2}\mathbf{x}^\mathsf{T}\left(\boldsymbol{\Sigma}_\mathbf{x}^{-1} + \boldsymbol{\Sigma}_\mathbf{z}^{-1}\right)\mathbf{x} + \frac{1}{2}\mathbf{x}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{y}^\mathsf{T}\boldsymbol{\Sigma}_\mathbf{z}^{-1}\mathbf{y} = -\frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{R}\mathbf{u}, \tag{2.204}$$

respectively, where

$$\mathbf{u} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \mathbf{v} = \begin{bmatrix} \boldsymbol{\Sigma}_\mathbf{x}^{-1}\boldsymbol{\mu}_\mathbf{x} - \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} \\ \boldsymbol{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \boldsymbol{\Sigma}_\mathbf{x}^{-1} & -\boldsymbol{\Sigma}_\mathbf{z}^{-1} \\ -\boldsymbol{\Sigma}_\mathbf{z}^{-1} & \boldsymbol{\Sigma}_\mathbf{z}^{-1} \end{bmatrix}.$$

Therefore, the logarithm of the integrand except the terms independent of $\mathbf{u}$ can be written as

$$-\frac{1}{2}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right)^{\mathsf{T}}\mathbf{R}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right), \tag{2.205}$$

where

$$\mathbf{R}^{-1}=\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}}+\boldsymbol{\Sigma}_{\mathbf{z}}\end{bmatrix}, \mathbf{R}^{-1}\mathbf{v}=\begin{bmatrix}\boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{x}}+\boldsymbol{\mu}_{\mathbf{z}}\end{bmatrix}.$$

by 2.29 and 2.30. Thus,

$$p(\mathbf{y})=\mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{x}}+\boldsymbol{\mu}_{\mathbf{z}},\boldsymbol{\Sigma}_{\mathbf{x}}+\boldsymbol{\Sigma}_{\mathbf{z}}\right). \tag{2.206}$$

2.32

Let $\mathbf{x}$ and $\mathbf{y}$ be variables such that

$$\begin{aligned}p(\mathbf{x})&=\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right), \\ p(\mathbf{y}|\mathbf{x})&=\mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1}\right).\end{aligned} \tag{2.207}$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x})=p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \tag{2.208}$$

The logarithm of the left hand side except the terms independent of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right)^{\mathsf{T}}\mathbf{L}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right)-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.209}$$

Since the first term can be written as

$$\begin{aligned}&-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right)^{\mathsf{T}}\mathbf{L}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right) \\ =&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})+(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}) \\ &-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\mathsf{T}}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}),\end{aligned} \tag{2.210}$$

the logarithm except the terms independent of $\mathbf{x}$ and $\mathbf{y}$ can be written as

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\mathsf{T} \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) + \frac{1}{2}\mathbf{z}^\mathsf{T}\left(\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} + \boldsymbol{\Lambda}\right)\mathbf{z} \\
&-\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\
={}&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\mathsf{T}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) \\
&-\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{M}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}),
\end{aligned}
\tag{2.211}
$$

where

$$
\begin{aligned}
\mathbf{z} &= \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}), \\
\mathbf{M} &= \mathbf{L} - \mathbf{L}\mathbf{A}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}.
\end{aligned}
\tag{2.212}
$$

we have

$$
\boldsymbol{\mu} + \mathbf{z} = \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right).
\tag{2.213}
$$

By 2.26,

$$
\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}.
\tag{2.214}
$$

Therefore,

$$
\mathbf{M} = \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}.
\tag{2.215}
$$

Thus,

$$
\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{x}|\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right), \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\right), \\
p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right).
\end{aligned}
\tag{2.216}
$$

2.33

Refer to 2.32.

2.34

Let $\mathbf{X}$ be a set of N variables such that

$$
\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).
\tag{2.217}
$$

To maximise it with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, setting the partial derivatives to zero gives

$$\mathbf{0} = \sum_{n=1}^{N} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1} \right)^{\mathsf{T}} \right) (\mathbf{x}_n - \boldsymbol{\mu}),$$

$$\mathbf{O} = -\frac{N}{2} \left(\boldsymbol{\Sigma}^{-1} \right)^{\mathsf{T}} + \frac{1}{2} \left(\boldsymbol{\Sigma}^{-1} \right)^2 \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.218}$$

Therefore,

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n,$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}. \tag{2.219}$$

2.35

Let $\mathbf{x}$ be a variable such that

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right). \tag{2.220}$$

Then

$$\mathbb{E}\,\mathbf{x}\mathbf{x}^{\mathsf{T}} = \int \mathbf{x}\mathbf{x}^{\mathsf{T}} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x}. \tag{2.221}$$

The right hand side can be written as

$$\int (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x}$$

$$= \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} + \boldsymbol{\mu} \int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} \tag{2.222}$$

$$+ \left(\int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} \right) \boldsymbol{\mu}^{\mathsf{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}} \int \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x}.$$

Since

$$\int \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} = 1,$$

$$\int \mathbf{x} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} = \boldsymbol{\mu}, \tag{2.223}$$

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) d\mathbf{x} = \boldsymbol{\Sigma},$$

the right hand side can be written as $\boldsymbol{\Sigma} + \boldsymbol{\mu\mu}^\mathsf{T}$. Therefore,

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^\mathsf{T} = \boldsymbol{\Sigma} + \boldsymbol{\mu\mu}^\mathsf{T}. \tag{2.224}$$

Additionally, let $\mathbf{x}_n$ and $\mathbf{x}_m$ be variables such that

$$\begin{aligned} p(\mathbf{x}_n) &= \mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \\ p(\mathbf{x}_m) &= \mathcal{N}\left(\mathbf{x_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \end{aligned} \tag{2.225}$$

If $n \neq m$, then

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^\mathsf{T} = \mathrm{E}\,\mathbf{x}_n\,\mathrm{E}\,\mathbf{x}_m^\mathsf{T}. \tag{2.226}$$

The right hand side can be written as $\boldsymbol{\mu\mu}^\mathsf{T}$. Therefore,

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^\mathsf{T} = \delta_{nm}\boldsymbol{\Sigma} + \boldsymbol{\mu\mu}^\mathsf{T}. \tag{2.227}$$

Finally, let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables such that

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.228}$$

By 2.34,

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}. \end{aligned} \tag{2.229}$$

Then

$$\mathrm{E}\,\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}. \tag{2.230}$$

The right hand side can be written as

$$\begin{aligned} &\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\,\mathbf{x}_n\mathbf{x}_n^\mathsf{T} - \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\mathbf{x}_n^\mathsf{T} - \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\,\mathbf{x}_n\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\mathsf{T} \\ &+ \frac{1}{N^3}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\mathsf{T}. \end{aligned} \tag{2.231}$$

The first term can be written as $\boldsymbol{\Sigma} + \boldsymbol{\mu\mu}^\mathsf{T}$. The second and third terms can be written as

$$-\frac{1}{N}\left((\boldsymbol{\Sigma} + \boldsymbol{\mu\mu}^\mathsf{T}) + (N-1)\boldsymbol{\mu\mu}^\mathsf{T}\right) = -\frac{1}{N}\boldsymbol{\Sigma} - \boldsymbol{\mu\mu}^\mathsf{T}. \tag{2.232}$$

The fourth term can be written as

$$\frac{1}{N^2}\left(N\left(\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right)+N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right)=\frac{1}{N}\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.233}$$

Therefore,

$$\mathrm{E}\,\mathbf{\Sigma}_{\mathrm{ML}}=\frac{N-1}{N}\mathbf{\Sigma}. \tag{2.234}$$

2.36

Let $x_1,\cdots,x_N$ be variables such that

$$p(x_n)=\mathcal{N}\left(x_n|\mu,\sigma^2\right). \tag{2.235}$$

Let us assume that μ is known. Then, by 2.34,

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2. \tag{2.236}$$

The right hand side can be written as

$$\frac{1}{N}(x_N-\mu)^2+\frac{1}{N}\sum_{n=1}^{N-1}(x_n-\mu)^2=\frac{1}{N}(x_N-\mu)^2+\frac{N-1}{N}\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}. \tag{2.237}$$

Therefore,

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)}=\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}+\frac{1}{N}\left((x_N-\mu)^2-\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}\right). \tag{2.238}$$

Since

$$\frac{\partial}{\partial\sigma^2}\left(-\ln p\left(x_n|\sigma^2\right)\right)=\frac{1}{2\sigma^2}-\frac{1}{2\left(\sigma^2\right)^2}(x_n-\mu)^2, \tag{2.239}$$

we have

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)}=\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}-\frac{\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}}{N}\frac{\partial}{\partial\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}}\left(-\ln p\left(x_N|\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}\right)\right). \tag{2.240}$$

2.37

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables such that

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.241}$$

Let us assume that $\boldsymbol{\mu}$ is known. Then, by 2.34,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.242}$$

The right hand side can be written as

$$\begin{aligned}
&\frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} + \frac{1}{N} \sum_{n=1}^{N-1} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}} \\
&= \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} + \frac{N-1}{N}\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}.
\end{aligned} \tag{2.243}$$

Therefore,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left((\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right). \tag{2.244}$$

Since

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(-\ln p(x_n|\boldsymbol{\Sigma})\right) = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2 (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}}, \tag{2.245}$$

we have

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} - \frac{\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}}{N} \frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}}\left(-\ln p\left(\mathbf{x}_N | \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)\right). \tag{2.246}$$

2.38

Let $x_1, \cdots, x_N$ be variables such that

$$\begin{aligned}
p(x_n|\mu) &= \mathcal{N}\left(x_n|\mu, \sigma^2\right), \\
p(\mu) &= \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).
\end{aligned} \tag{2.247}$$

By the Bayes' theorem,

$$p(\mu|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu)p(\mu). \tag{2.248}$$

The logarithm of the right hand side excpt the terms independent of $\mathbf{x}$ and μ can be written as

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2. \tag{2.249}$$

The first term can be written as

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\text{ML}}+\mu_{\text{ML}}-\mu)^2 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\text{ML}})^2 - \frac{N}{2\sigma^2}(\mu_{\text{ML}}-\mu)^2. \tag{2.250}$$

where

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n, \tag{2.251}$$

as derived in 2.34. Therefore, the logarithm except the terms independent of $\mathbf{x}$ and μ can be written as

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\text{ML}})^2 - \frac{N}{2\sigma^2}(\mu_{\text{ML}}-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$$
$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\text{ML}})^2 - \frac{1}{2\sigma_N^2}(\mu-\mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2}, \tag{2.252}$$

where

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0,$$
$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}. \tag{2.253}$$

Therefore,

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu \mid \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0, \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}\right). \tag{2.254}$$

2.39   (Incomplete)

Let $x_1,\cdots,x_N$ be variables such that

$$p(x_n|\mu) = \mathcal{N}\left(x_n|\mu,\sigma^2\right),$$
$$p(\mu) = \mathcal{N}\left(\mu|\mu_0,\sigma_0^2\right). \tag{2.255}$$

Then, by 2.38,
$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right), \tag{2.256}$$

where

$$\mu_N = \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^{N} x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0,$$
$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}. \tag{2.257}$$

Then

$$\mu_N = \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2}x_N + \frac{\sigma^2 - \sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_0,$$
$$\sigma_N^2 = \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\sigma_{N-1}^2. \tag{2.258}$$

Additionally, we have

$$p(\mu|x_1, \cdots, x_{N-1}) = \mathcal{N}\left(\mu|\mu_{N-1}, \sigma_{N-1}^2\right),$$
$$p(x_N|\mu) = \mathcal{N}\left(x_N|\mu, \sigma^2\right). \tag{2.259}$$

Then, $\ln p(\mu|x_1, \cdots, x_{N-1}) + \ln p(x_N|\mu)$ except the terms independent of μ or x_N can be written as

$$-\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2$$
$$= -\frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}}\left(\mu - \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} - \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N\right)^2$$
$$+ \frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}}\left(\frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N\right) - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \tag{2.260}$$

Thereofere,

$$\mu_N = \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N,$$
$$\sigma_N^2 = \frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}. \tag{2.261}$$

2.40

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables such that

$$p(\mathbf{x}_n|\boldsymbol{\mu}) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}\right). \tag{2.262}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \tag{2.263}$$

The logarithm of the right hand side excpt the terms independent of $\mathbf{X}$ and $\boldsymbol{\mu}$ can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathsf{T}}. \tag{2.264}$$

The first term can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$

$$= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}). \tag{2.265}$$

where

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{2.266}$$

as derived in 2.34. Therefore, the logarithm except the terms independent of $\mathbf{X}$ and $\boldsymbol{\mu}$ can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$

$$-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathsf{T}}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^{\mathsf{T}}\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)$$

$$+\frac{1}{2}\boldsymbol{\mu}_N^{\mathsf{T}}\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N, \tag{2.267}$$

where

$$\boldsymbol{\mu}_N = \left(N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right),$$
$$\boldsymbol{\Sigma}_N = \left(N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}. \tag{2.268}$$

Therefore,

$$p(\boldsymbol{\mu}|\mathbf{X}) = \mathcal{N}\left(\boldsymbol{\mu} \mid \left(N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right), \left(N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right). \tag{2.269}$$

2.41

By the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.270}$$

Then

$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a-1}\exp(-b\lambda)d\lambda. \tag{2.271}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.272}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1}\exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{\Gamma(a)}\int_0^\infty \lambda'^{a-1}\exp(-\lambda')d\lambda'. \tag{2.273}$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)}\Gamma(a) = 1. \tag{2.274}$$

Therefore,

$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = 1. \tag{2.275}$$

2.42

Let λ be a variable such that

$$p(\lambda) = \mathrm{Gam}(\lambda|a,b). \tag{2.276}$$

75

By the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.277}$$

Then

$$\mathrm{E}\,\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right)d\lambda. \tag{2.278}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.279}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b\Gamma(a)}\int_0^\infty \lambda'^a \exp(-\lambda')d\lambda'. \tag{2.280}$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)}\Gamma(a+1) = \frac{a}{b}. \tag{2.281}$$

Therefore,

$$\mathrm{E}\,\lambda = \frac{a}{b}. \tag{2.282}$$

Additionally,

$$\mathrm{E}\,\lambda^2 = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a+1}\exp\left(-\frac{\lambda}{b}\right)d\lambda. \tag{2.283}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.284}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b^2\Gamma(a)}\int_0^\infty \lambda'^{a+1}\exp(-\lambda')d\lambda'. \tag{2.285}$$

The right hand side can be written as

$$\frac{1}{b^2\Gamma(a)}\Gamma(a+2) = \frac{a(a+1)}{b^2}. \tag{2.286}$$

Therefore,
$$\mathrm{E}\,\lambda^2 = \frac{a(a+1)}{b^2}. \tag{2.287}$$

By the definition,
$$\mathrm{var}\,\lambda = \mathrm{E}\,\lambda^2 - (\mathrm{E}\,\lambda)^2. \tag{2.288}$$

Therefore,
$$\mathrm{var}\,\lambda = \frac{a}{b^2}. \tag{2.289}$$

Finally, setting the derivative of $\mathrm{Gam}(\lambda|a,b)$ with respect to λ to zero gives
$$0 = \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{\lambda} - b \right) \lambda^{a-1} \exp\left(-\frac{\lambda}{b} \right). \tag{2.290}$$

Therefore,
$$\mathrm{mode}\,\lambda = \frac{a-1}{b}. \tag{2.291}$$

2.43

Let
$$p\left(x|\sigma^2, q\right) = \frac{q}{2\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2} \right). \tag{2.292}$$

Then
$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = \frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2} \right) dx. \tag{2.293}$$

By the transformation
$$x' = \frac{x^q}{2\sigma^2}, \tag{2.294}$$

the right hand side can be written as
$$\begin{aligned}
&\frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') \left(2\sigma^2\right)^{\frac{1}{q}} \frac{1}{q} x'^{\frac{1}{q}-1} dx' \\
&= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x'^{\frac{1}{q}-1} \exp(-x') dx'.
\end{aligned} \tag{2.295}$$

The right hand side can be written as
$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q} \right) = 1. \tag{2.296}$$

Therefore,
$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = 1. \tag{2.297}$$

Additionally,
$$p\left(x|\sigma^2, 2\right) = \frac{1}{\Gamma(\frac{1}{2})} \left(2\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{2.298}$$

Therefore,
$$p\left(x|\sigma^2, 2\right) = \mathcal{N}\left(x|0, \sigma^2\right). \tag{2.299}$$

Finally, let $\mathbf{t} = (t_1, \cdots, t_N)^\mathsf{T}$ and $\mathbf{X} = \{x_1, \cdots, x_N\}$ such that
$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \tag{2.300}$$

where
$$p(\epsilon_n) = p\left(\epsilon_n|\sigma^2, q\right). \tag{2.301}$$

Therefore, the logarithm of $p(\epsilon_n)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as
$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q} \ln\left(2\sigma^2\right). \tag{2.302}$$

Thus, the logarithm of $p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2\right)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as
$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln\left(2\sigma^2\right). \tag{2.303}$$

2.44

Let $x_1, \cdots, x_N$ be variables such that
$$\begin{aligned} p(x_n|\mu, \tau) &= \mathcal{N}\left(x_n|\mu, \tau^{-1}\right), \\ p(\mu, \tau) &= \mathcal{N}\left(\mu|\mu_0, (\beta\tau)^{-1}\right) \mathrm{Gam}(\tau|a, b). \end{aligned} \tag{2.304}$$

By the Bayes' theorem,
$$p(\mu, \tau|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu, \tau)p(\mu, \tau). \tag{2.305}$$

The logrithm of the right hand side except the terms independent of $\mathbf{x}$, μ and τ can be written as

$$\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{2}\ln\tau - \frac{\beta\tau}{2}(\mu-\mu_0)^2 + (a-1)\ln\tau - b\tau$$

$$= \left(a + \frac{N-1}{2}\right)\ln\tau - \frac{N\tau}{2}(\bar{x}-\mu)^2 - \frac{\beta\tau}{2}(\mu-\mu_0)^2 - b\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2,$$

$$(2.306)$$

where

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{2.307}$$

Since

$$-\frac{N\tau}{2}(\bar{x}-\mu)^2 - \frac{\beta\tau}{2}(\mu-\mu_0)^2 = -\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x}+\beta\mu_0}{N+\beta}\right)^2 - \frac{N\beta\tau(\bar{x}-\mu_0)^2}{2(N+\beta)},$$

$$(2.308)$$

the right hand side can be written as

$$-\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x}+\beta\mu_0}{N+\beta}\right)^2$$

$$+ \left(a + \frac{N-1}{2}\right)\ln\tau - \left(b + \frac{N\beta(\bar{x}-\mu_0)^2}{2(N+\beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2\right)\tau.$$

$$(2.309)$$

Therefore,

$$p(\mu,\tau|\mathbf{x}) = \mathcal{N}\left(\mu \mid \frac{N\bar{x}+\beta\mu_0}{N+\beta}, ((N+\beta)\tau)^{-1}\right)$$

$$\text{Gam}\left(\tau \mid a + \frac{N+1}{2}, b + \frac{N\beta(\bar{x}-\mu_0)^2}{2(N+\beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2\right).$$

$$(2.310)$$

2.45

Let $\mathbf{x}$ be a variable in D dimensions such that

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right). \tag{2.311}$$

Then

$$p(\mathbf{X}|\mathbf{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}\right). \tag{2.312}$$

The right hand side exept the terms independent of $\mathbf{\Lambda}$ can be written as

$$(\det \mathbf{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})\right) = (\det \mathbf{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{S}\mathbf{\Lambda}\right)\right), \tag{2.313}$$

where

$$\mathbf{S} = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.314}$$

Therefore,

$$p(\mathbf{X}|\mathbf{\Lambda}) \propto (\det \mathbf{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{S}\mathbf{\Lambda}\right)\right). \tag{2.315}$$

Let us assume that a prior distribution of $\mathbf{\Lambda}$ is given by

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu)(\det \mathbf{\Lambda})^{\frac{\nu - D - 1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{\Lambda}\right)\right). \tag{2.316}$$

Then, by the definition,

$$p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) \propto p(\mathbf{X}|\mathbf{\Lambda})\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu), \tag{2.317}$$

where the right hand side except the terms independent of $\mathbf{\Lambda}$ can be written as

$$(\det \mathbf{\Lambda})^{\frac{\nu + N - D - 1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\left(\mathbf{W}^{-1} + \mathbf{S}\right)\mathbf{\Lambda}\right)\right). \tag{2.318}$$

Therefore,

$$p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) = \mathcal{W}\left(\mathbf{\Lambda} \mid \left(\mathbf{W}^{-1} + \mathbf{S}\right)^{-1}, \nu + N\right). \tag{2.319}$$

Thus, $\mathcal{W}$ is a conjugate prior distribution of $\mathbf{\Lambda}$.

2.46

Let x be a variable such that

$$p(x|\mu, \tau, a, b) = \mathcal{N}\left(x|\mu, \tau^{-1}\right)\operatorname{Gam}(\tau|a, b). \tag{2.320}$$

Then
$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}\left(x|\mu, \tau^{-1}\right) \text{Gam}(\tau|a, b) d\tau. \tag{2.321}$$

The right hand side can be written as
$$\int_0^\infty \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x - \mu)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau$$
$$= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{(x-\mu)^2}{2}\right)\tau\right) d\tau. \tag{2.322}$$

By the transformation
$$\tau' = \left(b + \frac{(x-\mu)^2}{2}\right)\tau, \tag{2.323}$$

the integral of the right hand side can be written as
$$\int_0^\infty \left(\frac{\tau'}{b + \frac{(x-\mu)^2}{2}}\right)^{a-\frac{1}{2}} \exp(-\tau') \frac{d\tau'}{b + \frac{(x-\mu)^2}{2}} = \Gamma\left(a + \frac{1}{2}\right)\left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.324}$$

Therefore,
$$p(x|\mu, \tau, a, b) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.325}$$

Let
$$\nu = 2a,$$
$$\lambda = \frac{a}{b}. \tag{2.326}$$

Then
$$p(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{2.327}$$

2.47

By the definition,
$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{2.328}$$

By the transformation

$$y = \frac{\lambda(x-\mu)^2}{\nu}, \tag{2.329}$$

the right hand side except the terms independent of x can be written as

$$(1+y)^{-\frac{\lambda(x-\mu)^2}{2y}-\frac{1}{2}}. \tag{2.330}$$

In the limit $y \to \infty$, it becomes

$$\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \tag{2.331}$$

Therefore, in the limit $\nu \to \infty$, $\mathrm{St}(x|\mu, \lambda, \nu)$ becomes $\mathcal{N}(x|\mu, \lambda^{-1})$.

2.48

Let $\mathbf{x}$ be a variable in D dimensions such that

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right). \tag{2.332}$$

Then

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.333}$$

The right hand side can be written as

$$\int_0^\infty (2\pi)^{-\frac{D}{2}} (\det(\eta\boldsymbol{\Lambda}))^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\eta\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right) \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \eta^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2}\eta\right) d\eta$$

$$= (2\pi)^{-\frac{D}{2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} (\det\boldsymbol{\Lambda})^{\frac{1}{2}} \int_0^\infty \eta^{\frac{D+\nu}{2}-1} \exp\left(-\frac{1}{2}\left(\nu + (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\eta\right) d\eta. \tag{2.334}$$

By the transformation

$$\eta' = \frac{1}{2}\left(\nu + (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\eta, \tag{2.335}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{2\eta'}{\nu + (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}-1} \exp(-\eta') \frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})} d\eta'$$

$$= \left(\frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}} \Gamma\left(\frac{D+\nu}{2}\right). \tag{2.336}$$

Therefore,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.337)$$

2.49

Let $\mathbf{x}$ be a variable such that

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu). \quad (2.338)$$

By the definition,

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.339)$$

First,

$$\mathrm{E}\,\mathbf{x} = \int \mathbf{x}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)d\mathbf{x}. \quad (2.340)$$

The right hand side can be written as

$$\int \mathbf{x} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta\right) d\mathbf{x}$$
$$= \int \left(\int \mathbf{x}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) d\mathbf{x}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.341)$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \quad (2.342)$$

Therefore,

$$\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu}. \quad (2.343)$$

Additionally,

$$\mathrm{cov}\,\mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)d\mathbf{x}. \quad (2.344)$$

The right hand side can be written as

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta\right) d\mathbf{x}$$
$$= \int \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) d\mathbf{x}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.345)$$

The right hand side can be written as

$$\int (\eta\mathbf{\Lambda})^{-1} \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \mathbf{\Lambda}^{-1} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \int \eta^{\frac{\nu}{2}-2} \exp\left(-\frac{\nu}{2}\eta\right) d\eta. \quad (2.346)$$

By the transformation

$$\eta' = \frac{\nu}{2}\eta, \quad (2.347)$$

the integral of the right hand side can be written as

$$\int \left(\frac{2}{\nu}\eta'\right)^{-\frac{\nu}{2}-2} \exp(-\eta')\frac{2}{\nu}d\eta' = \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right). \quad (2.348)$$

Therefore, the right hand side can be written as

$$\mathbf{\Lambda}^{-1} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right) = \frac{\frac{\nu}{2}}{\frac{\nu}{2}-1}\mathbf{\Lambda}^{-1}. \quad (2.349)$$

Thus,

$$\text{cov}\,\mathbf{x} = \frac{\nu}{\nu-2}\mathbf{\Lambda}^{-1}. \quad (2.350)$$

Finally, setting the derivative of $\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)$ with respect to $\mathbf{x}$ to zero gives

$$\mathbf{0} = -\frac{1}{2}\left(\mathbf{\Lambda} + \mathbf{\Lambda}^{\mathsf{T}}\right)(\mathbf{x} - \boldsymbol{\mu})\int \eta\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right)\text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right)d\eta. \quad (2.351)$$

Therefore,

$$\text{mode}\,\mathbf{x} = \boldsymbol{\mu}. \quad (2.352)$$

2.50

By the definition,

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}\frac{(\det\mathbf{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}}\left(1 + \frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.353)$$

By the transformation

$$y = \frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}, \quad (2.354)$$

the right hand side except the terms independent of $\mathbf{x}$ can be written as

$$(1 + y)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2y} - \frac{D}{2}}. \tag{2.355}$$

In the limit $y \to \infty$, it becomes

$$\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{2.356}$$

Therefore, in the limit $\nu \to \infty$, $\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$ becomes $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.

2.51

We have

$$\exp(iA)\exp(-iA) = 1. \tag{2.357}$$

The left hand side can be written as

$$(\cos A + i \sin A)(\cos A - i \sin A) = \cos^2 A + \sin^2 A. \tag{2.358}$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \tag{2.359}$$

Additionally,

$$\cos(A - B) = \mathrm{Re}\left(\exp\left(i(A - B)\right)\right). \tag{2.360}$$

The right hand side can be written as

$$\mathrm{Re}\left(\exp(iA)\exp(-iB)\right) = \mathrm{Re}\left((\cos A + i \sin A)(\cos B - i \sin B)\right). \tag{2.361}$$

The right hand side can be written as $\cos A \cos B + \sin A \sin B$. Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \tag{2.362}$$

Finally,

$$\sin(A - B) = \mathrm{Im}\left(\exp\left(i(A - B)\right)\right). \tag{2.363}$$

The right hand side can be written as

$$\mathrm{Im}\left(\exp(iA)\exp(-iB)\right) = \left((\cos A + i \sin A)(\cos B - i \sin B)\right). \tag{2.364}$$

The right hand side can be written as $\sin A \cos B - \cos A \sin B$. Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \tag{2.365}$$

2.52 (Incomplete)

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \tag{2.366}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta. \tag{2.367}$$

By the Taylor series

$$\cos\alpha = 1 - \frac{1}{2}\alpha^2 + O\left(\alpha^4\right) \tag{2.368}$$

and the transformation

$$\xi = m^{\frac{1}{2}}(\theta - \theta_0), \tag{2.369}$$

we have

$$\exp\left(m\cos(\theta - \theta_0)\right) = \exp\left(m\left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O\left((\theta - \theta_0)^4\right)\right)\right). \tag{2.370}$$

2.53

Let θ_0 be a parameter such that

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0. \tag{2.371}$$

The left hand side can be written as

$$\sum_{n=1}^{N} (\sin\theta_n \cos\theta_0 - \cos\theta_n \sin\theta_0) = \cos\theta_0 \sum_{n=1}^{N} \sin\theta_n - \sin\theta_0 \sum_{n=1}^{N} \cos\theta_n. \tag{2.372}$$

Therefore,

$$\theta_0 = \arctan\left(\frac{\sum_{n=1}^{N} \sin\theta_n}{\sum_{n=1}^{N} \cos\theta_n}\right). \tag{2.373}$$

2.54

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \qquad (2.374)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta. \qquad (2.375)$$

Setting the first and seconnd derivatives with respect to θ to zero gives

$$\begin{aligned} 0 &= -m\sin(\theta - \theta_0)p(\theta|\theta_0, m), \\ 0 &= \left(m^2\sin^2(\theta - \theta_0) - m\cos(\theta - \theta_0)\right)p(\theta|\theta_0, m). \end{aligned} \qquad (2.376)$$

Therefore,

$$\begin{aligned} \underset{\theta}{\operatorname{argmax}}\, p(\theta|\theta_0, m) &= \theta_0, \\ \underset{\theta}{\operatorname{argmin}}\, p(\theta|\theta_0, m) &= \theta_0 - \pi\operatorname{sgn}(\theta_0 - \pi). \end{aligned} \qquad (2.377)$$

2.55

Let

$$\theta_0^{\mathrm{ML}} = \arctan\left(\frac{\sum_{n=1}^N \sin\theta_n}{\sum_{n=1}^N \cos\theta_n}\right). \qquad (2.378)$$

Let

$$\begin{aligned} \bar{r}\cos\bar{\theta} &= \frac{1}{N}\sum_{n=1}^N \cos\theta_n, \\ \bar{r}\sin\bar{\theta} &= \frac{1}{N}\sum_{n=1}^N \sin\theta_n. \end{aligned} \qquad (2.379)$$

Then

$$\theta_0^{\mathrm{ML}} = \bar{\theta}. \qquad (2.380)$$

Here,

$$\frac{1}{N}\sum_{n=1}^N \cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \left(\frac{1}{N}\sum_{n=1}^N \cos\theta_n\right)\cos\theta_0^{\mathrm{ML}} + \left(\frac{1}{N}\sum_{n=1}^N \sin\theta_n\right)\sin\theta_0^{\mathrm{ML}}. \qquad (2.381)$$

By the result above, the right hand side can be written as

$$\bar{r}\cos^2\bar{\theta} + \bar{r}\sin^2\bar{\theta} = \bar{r}. \tag{2.382}$$

Therefore,

$$\frac{1}{N}\sum_{n=1}^{N}\cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \bar{r}. \tag{2.383}$$

2.56

By the definition,

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}. \tag{2.384}$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left((a-1)\ln\mu + (b-1)\ln(1-\mu)\right) \tag{2.385}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}.$$

Additionally, by the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.386}$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\exp\left((a-1)\ln\lambda - b\lambda\right). \tag{2.387}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}.$$

Finally, for

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)}\exp\left(m\cos(\theta-\theta_0)\right), \tag{2.388}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta, \tag{2.389}$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m\cos\theta_0\cos\theta + m\sin\theta_0\sin\theta). \tag{2.390}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{bmatrix}.$$

2.57

By the definition,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.391}$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right), \tag{2.392}$$

where

$$h(\mathbf{x}) = (2\pi)^{-\frac{D}{2}},$$

$$g(\boldsymbol{\eta}) = (\det(-2\boldsymbol{\eta}_2))^{-\frac{1}{2}} \exp\left(\frac{1}{4}\boldsymbol{\eta}_1^\mathsf{T}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1\right),$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix},$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^\mathsf{T} \end{bmatrix}.$$

2.58

Let $\mathbf{x}$ be a variable such that

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right). \tag{2.393}$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = 1 \tag{2.394}$$

with respect to $\boldsymbol{\eta}$ gives

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right) d\mathbf{x} + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right) d\mathbf{x} = \mathbf{0}.$$
(2.395)

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathrm{E}\,\mathbf{u}(\mathbf{x}).$$
(2.396)

Therefore,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}.$$
(2.397)

Thus,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}).$$
(2.398)

Taking the second derivative with respect to $\boldsymbol{\eta}$ gives

$$\nabla\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right) d\mathbf{x} + 2\nabla g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})^{\mathsf{T}} h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right) d\mathbf{x}$$
$$+ g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}} h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right) d\mathbf{x} = \mathbf{O}.$$
(2.399)

The left hand side can be written as

$$\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int \mathbf{u}(\mathbf{x})^{\mathsf{T}} p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}} p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x}$$
$$= \frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2\,\mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^{\mathsf{T}} + \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right).$$
(2.400)

Therefore,

$$\mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\mathsf{T}}}{g^2(\boldsymbol{\eta})}.$$
(2.401)

By the definition,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right) - \mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^{\mathsf{T}}.$$
(2.402)

Thus,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\mathsf{T}}}{g^2(\boldsymbol{\eta})}.$$
(2.403)

Hence,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = -\nabla\nabla \ln g(\boldsymbol{\eta}).$$
(2.404)

2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right). \tag{2.405}$$

Then

$$\int p(x|\sigma)dx = \frac{1}{\sigma} \int f\left(\frac{x}{\sigma}\right) dx. \tag{2.406}$$

By the transformation

$$x' = \frac{x}{\sigma}, \tag{2.407}$$

the right hand side can be written as

$$\frac{1}{\sigma} \int f(x')\sigma dx' = \int f(x')dx'. \tag{2.408}$$

Therefore, $p(x|\sigma)$ will be normalised if $f(x)$ is normalised.

2.60

Let $\mathbf{x}$ be a variable such that

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \tag{2.409}$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \tag{2.410}$$

Since

$$\int p(\mathbf{x})d\mathbf{x} = 1, \tag{2.411}$$

we have

$$\sum_i h_i \Delta_i = 1. \tag{2.412}$$

Let N be the total number of observations and n_i be the number of observations which fall in $\mathcal{R}_i$. Then, the logarithm of the likelihood is given by

$$\ln\left(\prod_i h_i{}^{n_i}\right) = \sum_i n_i \ln h_i, \tag{2.413}$$

where
$$\sum_i n_i = N. \tag{2.414}$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda \left(\sum_i h_i \Delta_i - 1 \right) \tag{2.415}$$

with respect to h_i and λ to zero gives

$$\frac{n_i}{h_i} + \lambda \Delta_i = 0,$$
$$\sum_i h_i \Delta_i - 1 = 0. \tag{2.416}$$

Therefore,

$$\lambda = -N,$$
$$h_i = \frac{n_i}{N \Delta_i}. \tag{2.417}$$

Thus, the maximum likelihood estimator for the $\{h_i\}$ is $\frac{n_i}{N\Delta_i}$.

2.61   (Incomplete)

Let $\mathbf{x}$ be a variable and $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be observations. Let

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})}, \tag{2.418}$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}'-\mathbf{x}\| \leq \|\mathbf{x}_{(K)}-\mathbf{x}\|} d\mathbf{x}', \tag{2.419}$$

K is a constant and $\mathbf{x}_{(K)}$ is the Kth nearest observation from the point $\mathbf{x}$.

3   Linear Models for Regression

3.1

By the definition,
$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{3.1}$$

The right hand side can be written as
$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \tag{3.2}$$

Therefore,
$$\tanh a = 2\sigma(2a) - 1, \tag{3.3}$$

where
$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{3.4}$$

Let
$$y(x_n, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right). \tag{3.5}$$

By the result above, the right hand side can be written as

$$w_0 + \sum_{m=1}^{M} w_m \frac{1 + \tanh\left(\frac{x-\mu_m}{2s}\right)}{2} = w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m + \frac{1}{2}\sum_{m=1}^{M} w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \tag{3.6}$$

Therefore, $y(x_n, \mathbf{w})$ is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{m=1}^{M} u_m \tanh\left(\frac{x - \mu_m}{2s}\right), \tag{3.7}$$

where

$$u_0 = w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m,$$
$$u_m = \frac{1}{2}w_m. \tag{3.8}$$

3.2   (Incomplete)

Let $\mathbf{\Phi}$ be an $N \times M$ matarix. Then, for any vector $\mathbf{v}$ in N dimensions,

$$\mathbf{\Phi} \left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathsf{T}}\mathbf{v} \tag{3.9}$$

is a projection of $\mathbf{v}$ onto the space spanned by the columns of $\mathbf{\Phi}$?

Additionally, for a vector $\mathbf{t}$ in N dimensions,

$$\left(\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathsf{T}}\mathbf{t} \tag{3.10}$$

is an orthogonal projection of $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{\Phi}$?

3.3

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(t_n - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 . \tag{3.11}$$

The right hand side can be written as

$$\frac{1}{2}\|\mathbf{t}' - \mathbf{\Phi}'\mathbf{w}\|^2, \tag{3.12}$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1}t_1 \\ \vdots \\ \sqrt{r_N}t_N \end{bmatrix}, \mathbf{\Phi}' = \begin{bmatrix} \sqrt{r_1}\boldsymbol{\phi}(\mathbf{x}_1)^{\mathsf{T}} \\ \vdots \\ \sqrt{r_N}\boldsymbol{\phi}(\mathbf{x}_N)^{\mathsf{T}} \end{bmatrix}.$$

Setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = -\mathbf{\Phi}'^{\mathsf{T}}(\mathbf{t}' - \mathbf{\Phi}'\mathbf{w}). \tag{3.13}$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}) = \left(\mathbf{\Phi}'^{\mathsf{T}}\mathbf{\Phi}'\right)^{-1} \mathbf{\Phi}'^{\mathsf{T}}\mathbf{t}'. \tag{3.14}$$

3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right)^2, \tag{3.15}$$

where

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_m (x_m + \epsilon_m), \tag{3.16}$$

$$p(\epsilon_m) = \mathcal{N}\left(\epsilon_m | 0, \sigma^2 \right).$$

Setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right).$$

The right hand side can be written as

$$\sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right) + \sum_{n=1}^{N} \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right).$$

3.5

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n) \right)^2. \tag{3.17}$$

Then, the minimisation of $E(\mathbf{w})$ under the constraint

$$\sum_{m=1}^{M} |w_m|^q \leq \eta \tag{3.18}$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \tag{3.19}$$

with respect to $\mathbf{w}$ and λ. Then,

$$\eta = \sum_{m=1}^{M} |w_m^*(\lambda)|^q, \tag{3.20}$$

where

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \left(E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \right). \tag{3.21}$$

3.6

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in D dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}, \mathbf{\Sigma}) = \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{W}), \mathbf{\Sigma}\right), \tag{3.22}$$

where

$$\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n). \tag{3.23}$$

Then,

$$\ln\left(\prod_{n=1}^{N} p(\mathbf{t}_n | \mathbf{W}, \mathbf{\Sigma})\right)$$

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det \mathbf{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n)). \tag{3.24}$$

By 3.21(a), setting the derivatives with respect to $\mathbf{W}$ and $\mathbf{\Sigma}$ to zero gives

$$\mathbf{O} = -\frac{1}{2}\left(\mathbf{\Sigma}^{-1} + \left(\mathbf{\Sigma}^{-1}\right)^\mathsf{T}\right)\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))\left(\boldsymbol{\phi}(\mathbf{x}_n)\right)^\mathsf{T},$$

$$\mathbf{O} = -\frac{N}{2}\left(\mathbf{\Sigma}^{-1}\right)^\mathsf{T} + \frac{1}{2}\left(\mathbf{\Sigma}^{-1}\right)^2 \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T}. \tag{3.25}$$

Therefore,

$$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\mathsf{T}\mathbf{t},$$

$$\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T}, \tag{3.26}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}.$$

96

3.7

Let $t_1, \cdots, t_N$ be variables such that

$$
p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),
$$
$$
p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.27}
$$

By the Bayes' theorem,

$$
p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.28}
$$

The logarithm of the right hand side except the terms independent of $\mathbf{t}$ and $\mathbf{w}$ can be written as

$$
\begin{aligned}
&-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n))^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\
&= -\frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0),
\end{aligned} \tag{3.29}
$$

where

$$
\mathbf{\Phi} = \begin{bmatrix} \phi(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \phi(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}.
$$

The right hand side can be written as

$$
-\frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\mathsf{T}\mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right) + \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\mathbf{t}^\mathsf{T}\mathbf{t} - \frac{1}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0, \tag{3.30}
$$

where

$$
\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{t}\right),
$$
$$
\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}. \tag{3.31}
$$

Therefore,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \tag{3.32}
$$

3.8

Let $t_1, \cdots, t_N$ be variables such that

$$
p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),
$$
$$
p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.33}
$$

Then, by 3.7,
$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.34}$$

where
$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}_N^\intercal \mathbf{t}_N \right), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}_N^\intercal \boldsymbol{\Phi}_N.
\end{aligned} \tag{3.35}$$

By the Bayes' theorem,
$$p(\mathbf{w}|\mathbf{t}_{N+1}) p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w}) p(\mathbf{w}). \tag{3.36}$$

The right hand side can be written as
$$p(t_{N+1}|\mathbf{w}) p(\mathbf{t}_N|\mathbf{w}) p(\mathbf{w}) = p(t_{N+1}|\mathbf{w}) p(\mathbf{w}|\mathbf{t}_N) p(\mathbf{t}_N). \tag{3.37}$$

Therefore,
$$p(\mathbf{w}|\mathbf{t}_{N+1}) p(t_{N+1}) = p(t_{N+1}|\mathbf{w}) p(\mathbf{w}|\mathbf{t}_N). \tag{3.38}$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as

$$\begin{aligned}
&-\frac{\beta}{2} \left(t_{N+1} - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_{N+1}) \right)^2 - \frac{1}{2} \left(\mathbf{w} - \mathbf{m}_N \right)^\intercal \mathbf{S}_N^{-1} \left(\mathbf{w} - \mathbf{m}_N \right) \\
&= -\frac{1}{2} \left(\mathbf{w} - \boldsymbol{\mu}_{N+1} \right)^\intercal \boldsymbol{\Lambda}_{N+1} \left(\mathbf{w} - \boldsymbol{\mu}_{N+1} \right) + \frac{1}{2} \boldsymbol{\mu}_{N+1}^\intercal \boldsymbol{\Lambda}_{N+1} \boldsymbol{\mu}_{N+1} \\
&\quad - \frac{1}{2} \mathbf{m}_N^\intercal \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} t_{N+1}^2,
\end{aligned} \tag{3.39}$$

where
$$\begin{aligned}
\boldsymbol{\mu}_{N+1} &= \boldsymbol{\Lambda}_{N+1}^{-1} \left(\mathbf{S}_N^{-1} \mathbf{m}_N + \beta t_{N+1} \boldsymbol{\phi}(\mathbf{x}_{N+1}) \right), \\
\boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal.
\end{aligned} \tag{3.40}$$

Therefore,
$$\begin{aligned}
\boldsymbol{\mu}_{N+1} &= \mathbf{m}_{N+1}, \\
\boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_{N+1}^{-1}.
\end{aligned} \tag{3.41}$$

Thus,
$$p(\mathbf{w}|\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}). \tag{3.42}$$

3.9 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.43}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.44}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\mathsf{T}\mathbf{t}_N\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\mathsf{T}\boldsymbol{\Phi}_N. \tag{3.45}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \tag{3.46}$$

The right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \tag{3.47}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N). \tag{3.48}$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2 - \frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\mathsf{T}\mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right). \tag{3.49}$$

3.10

Let t be a variable such that

$$p(t|\mathbf{w}) = \mathcal{N}\left(t|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.50}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.51}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \tag{3.52}$$

By marginalisation,

$$p(t|\mathbf{t}) = \int p(t|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{3.53}$$

The logarithm of the integrand of the right hand side except the terms independent of t and $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\left(t - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x})\right)^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.54}$$

It can be written as

$$-\frac{1}{2}\begin{bmatrix}\mathbf{w}\\t\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\beta\boldsymbol{\phi}(\mathbf{x})\\-\beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}\begin{bmatrix}\mathbf{w}\\t\end{bmatrix} + \begin{bmatrix}\mathbf{w}\\t\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\0\end{bmatrix} - \frac{1}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\beta\boldsymbol{\phi}(\mathbf{x})\\-\beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{S}_N & \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\\\boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N & \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Therefore,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\boldsymbol{\phi}(\mathbf{x})\\\beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\0\end{bmatrix} = \begin{bmatrix}\mathbf{m}_N\\\mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Thus,

$$p(t|\mathbf{t}) = \mathcal{N}\left(t|\mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right), \tag{3.55}$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}). \tag{3.56}$$

3.11

Let t be a variable such that

$$\begin{aligned}p(t|\mathbf{w}) &= \mathcal{N}\left(t|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right),\\p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).\end{aligned} \tag{3.57}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.58}$$

where
$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{\Phi}_N^\mathsf{T} \mathbf{t}_N \right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}_N^\mathsf{T} \mathbf{\Phi}_N. \tag{3.59}$$

Then, by 3.10,
$$p(t|\mathbf{t}) = \mathcal{N}\left(t \mid \mathbf{m}_N^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}) \right), \tag{3.60}$$

where
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \tag{3.61}$$

Then,
$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \left(\mathbf{S}_N - \mathbf{S}_{N+1} \right) \boldsymbol{\phi}(\mathbf{x}). \tag{3.62}$$

By the expression of $\mathbf{S}_N$ above,
$$\mathbf{S}_{N+1} = \left(\mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T} \right)^{-1}. \tag{3.63}$$

By the identity
$$(\mathbf{M} + \mathbf{v}\mathbf{v}^\mathsf{T})^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})(\mathbf{v}^\mathsf{T}\mathbf{M}^{-1})}{1 + \mathbf{v}^\mathsf{T}\mathbf{M}^{-1}\mathbf{v}}, \tag{3.64}$$

the right hand side can be written as
$$\mathbf{S}_N - \frac{\beta \left(\mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1}) \right) \left(\boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T} \mathbf{S}_N \right)}{1 + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T}}. \tag{3.65}$$

Therefore,
$$\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}(\mathbf{S}_N - \mathbf{S}_{N+1})\boldsymbol{\phi}(\mathbf{x}) = \frac{\beta \left(\boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1}) \right)^2}{1 + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T}}. \tag{3.66}$$

Thus,
$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}). \tag{3.67}$$

3.12

Let $t_1, \cdots, t_N$ be variables such that
$$p(t_n|\mathbf{w}, \beta) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1} \right),$$
$$p(\mathbf{w}, \beta) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0 \right) \mathrm{Gam}(\beta|a_0, b_0), \tag{3.68}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{w}, \beta | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta). \tag{3.69}$$

The logarithm of the right hand side except the terms independent of $\mathbf{t}$, $\mathbf{w}$ and β can be written as

$$-\frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{M}{2} \ln \beta^{-1} - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^{\mathsf{T}} \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)$$

$$+ (a_0 - 1) \ln \beta - b_0 \beta$$

$$= -\frac{M}{2} \ln \beta - \frac{\beta}{2} \mathbf{w}^{\mathsf{T}} \left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi} \right) \mathbf{w} + \beta \mathbf{w}^{\mathsf{T}} \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{t} \right) - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^{\mathsf{T}} \mathbf{S}_0^{-1} \mathbf{m}_0$$

$$+ \left(a_0 + \frac{N}{2} - 1 \right) \ln \beta - b_0 \beta. \tag{3.70}$$

The right hand side can be written as

$$-\frac{M}{2} \ln \beta - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^{\mathsf{T}} \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) + (a_N - 1) \ln \beta - b_N \beta, \tag{3.71}$$

where

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{t} \right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi},$$
$$a_N = a_0 + \frac{N}{2}, \tag{3.72}$$
$$b_N = b_0 + \frac{1}{2} \|\mathbf{t}\|^2 + \frac{1}{2} \mathbf{m}_0^{\mathsf{T}} \mathbf{S}_0 \mathbf{m}_0 - \frac{1}{2} \mathbf{m}_N^{\mathsf{T}} \mathbf{S}_N^{-1} \mathbf{m}_N.$$

Therefore,

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N} \left(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N \right) \mathrm{Gam}(\beta | a_N, b_N). \tag{3.73}$$

Substituting it to the result of the Bayes' theorem above, we have

$$p(\mathbf{t}) = \frac{\mathcal{N} \left(\mathbf{t} | \boldsymbol{\Phi} \mathbf{w}, \beta^{-1} \mathbf{I} \right) \mathcal{N} \left(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0 \right) \mathrm{Gam}(\beta | a_0, b_0)}{\mathcal{N} \left(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N \right) \mathrm{Gam}(\beta | a_N, b_N)}. \tag{3.74}$$

The logarithm of the right hand side can be written as

$$-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^{\mathsf{T}}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})$$

$$-\frac{M}{2}\ln(2\pi) - \frac{M}{2}\ln\beta^{-1} - \frac{1}{2}\det\mathbf{S}_0 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)$$

$$+ a_0\ln b_0 - \ln\Gamma(a_0) + (a_0 - 1)\ln\beta - b_0\beta$$

$$+\frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln\beta^{-1} + \frac{1}{2}\det\mathbf{S}_N + \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^{\mathsf{T}}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_0)$$

$$- a_N\ln b_N + \ln\Gamma(a_N) - (a_N - 1)\ln\beta + b_N\beta$$

$$= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\det\mathbf{S}_0 + a_0\ln b_0 - \ln\Gamma(a_0) + \frac{1}{2}\det\mathbf{S}_N - a_N\ln b_N + \ln\Gamma(a_N). \tag{3.75}$$

Therefore,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}}\left(\frac{\det\mathbf{S}_N}{\det\mathbf{S}_0}\right)^{\frac{1}{2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}. \tag{3.76}$$

3.13

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}, \beta) = \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}, \beta) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0), \tag{3.77}$$

where $\mathbf{w}$ and ϕ are vectors in M dimensions. Then, by 3.12,

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N), \tag{3.78}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^{\mathsf{T}}\mathbf{t}\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi},$$
$$a_N = a_0 + \frac{N}{2}, \tag{3.79}$$
$$b_N = b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0\mathbf{m}_0 - \frac{1}{2}\mathbf{m}_N^{\mathsf{T}}\mathbf{S}_N^{-1}\mathbf{m}_N.$$

By marginalisation,

$$p(t|\mathbf{t}) = \int\int p(t|\mathbf{w}, \beta)p(\mathbf{w}, \beta|\mathbf{t})d\mathbf{w}d\beta. \tag{3.80}$$

103

The right hand side can be written as

$$\int \left(\int \mathcal{N}\left(t|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right) d\mathbf{w} \right) \mathrm{Gam}(\beta|a_N, b_N) d\beta.$$

(3.81)

The logarithm of the integrand with respect to $\mathbf{w}$ except the terms indepndent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\left(t - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x})\right)^2 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N).$$

(3.82)

It can be written as

$$-\frac{\beta}{2}\begin{bmatrix}\mathbf{w}\\t\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\-\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}\begin{bmatrix}\mathbf{w}\\t\end{bmatrix} + \beta\begin{bmatrix}\mathbf{w}\\t\end{bmatrix}^\intercal\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\0\end{bmatrix} - \frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\-\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{S}_N & \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\\\boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N & 1 + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Then,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\-\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\0\end{bmatrix} = \begin{bmatrix}\mathbf{m}_N\\\mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Therefore, the integral with respect to $\mathbf{w}$ can be written as

$$\mathcal{N}\left(t|\mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)\right).$$

(3.83)

Then, the logarithm of the integrand with respect to β except the terms independent of β can be written as

$$-\frac{1}{2}\ln\beta^{-1} - \frac{\beta}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\left(t - \mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\right)^2 + (a_N - 1)\ln\beta - b_N\beta$$

$$= \left(a_N + \frac{1}{2} - 1\right)\ln\beta - \left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\right)\beta.$$

(3.84)

Therefore, the integral with respect to β except the terms independent of t can be written as

$$\left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\right)^{-a_N - \frac{1}{2}}.$$

(3.85)

Thus,
$$p(t|\mathbf{x}, \mathbf{t}) = \mathrm{St}(t|\mu, \lambda, \nu), \tag{3.86}$$

where
$$\begin{aligned}
\mu &= \mathbf{m}_N^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}), \\
\lambda &= \frac{a_N}{b_N} \left(1 + \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)^{-1}, \\
\nu &= 2a_N.
\end{aligned} \tag{3.87}$$

3.14 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that
$$\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),
\end{aligned} \tag{3.88}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. Then, by 3.7,
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.89}$$

where
$$\begin{aligned}
\mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\mathsf{T} \mathbf{t}, \\
\mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi}.
\end{aligned} \tag{3.90}$$

Let
$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}). \tag{3.91}$$

Then,
$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n, \tag{3.92}$$

where
$$k(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}'). \tag{3.93}$$

Let us suppose that $\phi_j(\mathbf{x})$ are linearly independent, $N > M$ and
$$\phi_0(\mathbf{x}) = 1. \tag{3.94}$$

Then, we can construct a new basis set $\psi_j(\mathbf{x})$ such that
$$\boldsymbol{\Psi}^\mathsf{T} \boldsymbol{\Psi} = \mathbf{I}? \tag{3.95}$$

$$\sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_k(\mathbf{x}_n) = I_{jk}? \tag{3.96}$$

where

$$\Psi = \begin{bmatrix} \boldsymbol{\psi}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\psi}(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}$$

and

$$\psi_0(\mathbf{x}) = 1. \tag{3.97}$$

Under the basis set, if $\alpha = 0$, then

$$\mathbf{S}_N^{-1} = \beta\mathbf{I}, \tag{3.98}$$

so that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\mathsf{T}\boldsymbol{\psi}(\mathbf{x}'). \tag{3.99}$$

Then,

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N}\sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) = 1? \tag{3.100}$$

3.15

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.101}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.102}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \end{aligned} \tag{3.103}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.104}$$

where
$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \tag{3.105}$$

By 3.22, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives
$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N},$$
$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}, \tag{3.106}$$

where
$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m} \tag{3.107}$$

and $\lambda_1, \cdots, \lambda_M$ are the eigenvalues of $\beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$. If α and β are set as above, then
$$E(\mathbf{m}_N) = \frac{N}{2}. \tag{3.108}$$

3.16

Let $t_1, \cdots, t_N$ be variables such that
$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.109}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.110}$$

Integrating both sides with respect to $\mathbf{w}$ gives
$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.111}$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be written as
$$-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} = -\frac{1}{2}\begin{bmatrix}\mathbf{w}\\\mathbf{t}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\alpha\mathbf{I} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} & -\beta\mathbf{\Phi}^\mathsf{T}\\-\beta\mathbf{\Phi} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\\mathbf{t}\end{bmatrix}.$$

By 2.24,

$$
\begin{bmatrix} \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathsf{T}} \mathbf{\Phi} & -\beta \mathbf{\Phi}^{\mathsf{T}} \\ -\beta \mathbf{\Phi} & \beta \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \alpha^{-1} \mathbf{I} & \alpha^{-1} \mathbf{\Phi}^{\mathsf{T}} \\ \alpha^{-1} \mathbf{\Phi} & \alpha^{-1} \mathbf{\Phi} \mathbf{\Phi}^{\mathsf{T}} + \beta^{-1} \mathbf{I} \end{bmatrix} .
$$

Therefore,
$$
p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t} | \mathbf{0}, \alpha^{-1} \mathbf{\Phi} \mathbf{\Phi}^{\mathsf{T}} + \beta^{-1} \mathbf{I}\right). \tag{3.112}
$$

3.17

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n | \mathbf{w}) &= \mathcal{N}\left(t_n | \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}),
\end{aligned} \tag{3.113}
$$

where $\mathbf{w}$ and ϕ are vectors in M dimensions. By the Bayes' theorem,

$$
p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \tag{3.114}
$$

Then,
$$
p(\mathbf{t}) = \int p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \tag{3.115}
$$

The logarithm of the integrand of the right hand side can be written as

$$
-\frac{N}{2} \ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n))^2 - \frac{M}{2} \ln\left(2\pi\right) - \frac{1}{2} \ln\left(\det\left(\alpha^{-1} \mathbf{I}\right)\right) - \frac{\alpha}{2} \mathbf{w}^{\mathsf{T}} \mathbf{w}. \tag{3.116}
$$

Therefore,
$$
p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w}, \tag{3.117}
$$

where
$$
E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^{\mathsf{T}} \mathbf{w}. \tag{3.118}
$$

3.18

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n | \mathbf{w}) &= \mathcal{N}\left(t_n | \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}),
\end{aligned} \tag{3.119}
$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.120}$$

where

$$\begin{aligned}\mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}.\end{aligned} \tag{3.121}$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int \exp\left(-E(\mathbf{w})\right)d\mathbf{w}, \tag{3.122}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{3.123}$$

The first term of the definition of $E(\mathbf{w})$ can be written as

$$\begin{aligned}&\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N - \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N)\|^2 \\ =&\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 - \beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N).\end{aligned} \tag{3.124}$$

Similarly, the second term can be written as

$$\begin{aligned}&\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N) \\ =&\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) + \alpha\mathbf{m}_N^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N.\end{aligned} \tag{3.125}$$

Here,

$$\begin{aligned}&-\beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \alpha\mathbf{m}_N^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) \\ =&\left(-\beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\mathbf{m}_N + \alpha\mathbf{m}_N\right)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N).\end{aligned} \tag{3.126}$$

By the definitions of $\mathbf{m}_N$ and $\mathbf{S}_N$ above, the right hand can be written as

$$\left(-\beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t} + \mathbf{S}_N^{-1}\mathbf{m}_N\right)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) = 0. \tag{3.127}$$

Therefore,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.128}$$

3.19

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),$$

(3.129)

where $\mathbf{w}$ and ϕ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N),$$

(3.130)

where

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^\mathsf{T}\mathbf{t},$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}.$$

(3.131)

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int \exp\left(-E(\mathbf{w})\right) d\mathbf{w},$$

(3.132)

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}.$$

(3.133)

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N).$$

(3.134)

Therefore, the integral in the expression above of $p(\mathbf{t})$ can be written as

$$\exp\left(-E(\mathbf{m}_N)\right)\int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N)\right) d\mathbf{w}$$
$$= (2\pi)^{\frac{M}{2}}(\det \mathbf{S}_N)^{\frac{1}{2}}\exp\left(-E(\mathbf{m}_N)\right).$$

(3.135)

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln \alpha + \frac{N}{2}\ln \beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}_N) - E(\mathbf{m}_N). \quad (3.136)$$

3.20

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),$$

(3.137)

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.138}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\mathsf{T} \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi}. \end{aligned} \tag{3.139}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.140}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\mathsf{T} \mathbf{m}_N. \tag{3.141}$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors of $\beta \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi}$ such that

$$\beta \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi} \mathbf{u}_m = \lambda_m \mathbf{u}_m. \tag{3.142}$$

Then,

$$\mathbf{S}_N^{-1} \mathbf{u}_m = (\alpha + \lambda_m) \mathbf{u}_m, \tag{3.143}$$

so that

$$\det \mathbf{S}_N = \prod_{m=1}^{M} \frac{1}{\alpha + \lambda_m}. \tag{3.144}$$

Therefore, setting the derivative of $\ln p(\mathbf{t}|\alpha, \beta)$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{m=1}^{M} \frac{1}{\alpha + \lambda_m} - \frac{1}{2} \mathbf{m}_N^\mathsf{T} \mathbf{m}_N. \tag{3.145}$$

Multiplying both sides by 2α gives

$$\alpha \mathbf{m}_N^\mathsf{T} \mathbf{m}_N = M - \sum_{m=1}^{M} \frac{\alpha}{\alpha + \lambda_m}. \tag{3.146}$$

The right hand side can be written as

$$\sum_{m=1}^{M} \left(1 - \frac{\alpha}{\alpha + \lambda_m} \right) = \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m}. \tag{3.147}$$

111

Thus,
$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N}, \tag{3.148}$$

where
$$\gamma = \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m}. \tag{3.149}$$

3.21

(a)

Let $\boldsymbol{\Sigma}$ be a $M \times M$ real symmetric matrix such that
$$\boldsymbol{\Sigma}\mathbf{u}_m = \lambda_m\mathbf{u}_m, \tag{3.150}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_M$ are unit vectors. Let
$$\begin{aligned}\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_M),\\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_M].\end{aligned} \tag{3.151}$$

By 2.19,
$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T},\\ \mathbf{U}^\mathsf{T}\mathbf{U} &= \mathbf{I}.\end{aligned} \tag{3.152}$$

Therefore,
$$\det\boldsymbol{\Sigma} = \prod_{m=1}^{M} \lambda_m, \tag{3.153}$$

so that
$$\ln(\det\boldsymbol{\Sigma}) = \sum_{m=1}^{M} \ln\lambda_i. \tag{3.154}$$

Then,
$$\frac{\partial}{\partial\alpha}\ln(\det\boldsymbol{\Sigma}) = \sum_{m=1}^{M} \frac{\partial\lambda_m}{\partial\alpha}\frac{1}{\lambda_m}. \tag{3.155}$$

Thus,
$$\frac{\partial}{\partial\alpha}\ln(\det\boldsymbol{\Sigma}) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial}{\partial\alpha}\boldsymbol{\Sigma}\right). \tag{3.156}$$

(b)

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.157}$$

where $\mathbf{w}$ and ϕ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.158}$$

where

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^\mathsf{T}\mathbf{t},$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}. \tag{3.159}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.160)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \tag{3.161}$$

By 3.21(a),

$$\frac{\partial}{\partial\alpha}\ln\left(\det\mathbf{S}_N^{-1}\right) = \operatorname{tr}\left(\mathbf{S}_N\right). \tag{3.162}$$

The right hand side can be written as

$$\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m}, \tag{3.163}$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m} - \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N, \tag{3.164}$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N}, \tag{3.165}$$

where

$$\gamma = \sum_{m=1}^{M}\frac{\lambda_m}{\alpha + \lambda_m}. \tag{3.166}$$

113

3.22

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),
\end{aligned}
\tag{3.167}
$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.168}
$$

where

$$
\begin{aligned}
\mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\
\mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}.
\end{aligned}
\tag{3.169}
$$

By 3.19,

$$
\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.170}
$$

where

$$
E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \tag{3.171}
$$

By 3.21(a),

$$
\frac{\partial}{\partial\beta}\ln\left(\det\mathbf{S}_N^{-1}\right) = \mathrm{tr}\left(\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right). \tag{3.172}
$$

Since

$$
\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} = \frac{1}{\beta}\left(\mathbf{I} - \alpha\mathbf{S}_N\right), \tag{3.173}
$$

the right hand side can be written as

$$
\frac{1}{\beta}\left(M - \alpha\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m}\right) = \frac{1}{\beta}\sum_{m=1}^{M}\frac{\lambda_m}{\alpha + \lambda_m}, \tag{3.174}
$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to β to zero gives

$$
0 = \frac{N}{2\beta} - \frac{1}{2\beta}\sum_{m=1}^{M}\frac{\lambda_i}{\alpha + \lambda_m} - \frac{1}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2. \tag{3.175}
$$

Thus,
$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}, \tag{3.176}$$
where
$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m}. \tag{3.177}$$

3.23

Let $t_1, \cdots, t_N$ be variables such that
$$\begin{aligned} p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right) \mathrm{Gam}(\beta|a_0, b_0), \end{aligned} \tag{3.178}$$
where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By marginalisation,
$$p(\mathbf{t}) = \int \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}, \beta)d\mathbf{w}d\beta. \tag{3.179}$$
The right hand side can be written as
$$\int \left(\int \left(\prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right) \right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right) d\mathbf{w} \right) \mathrm{Gam}(\beta|a_0, b_0)d\beta. \tag{3.180}$$
The logarithm of the integrand with respect to $\mathbf{w}$ can be written as
$$\begin{aligned} &-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 \\ &-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\beta^{-1}\mathbf{S}_0\right) - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\ =&-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\ &-\frac{\beta}{2}\mathbf{w}^\mathsf{T}\left(\mathbf{S}_0^{-1} + \mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^\mathsf{T}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^\mathsf{T}\mathbf{t}\right) - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0. \end{aligned} \tag{3.181}$$
The right hand side can be written as
$$\begin{aligned} &-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\ &-\frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0, \end{aligned} \tag{3.182}$$

where

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \mathbf{\Phi}^\intercal \mathbf{t} \right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \mathbf{\Phi}^\intercal \mathbf{\Phi}. \tag{3.183}$$

Therefore, the logarithm of the integral with respect to $\mathbf{w}$ can be written as

$$-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N)$$
$$+ \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal \mathbf{S}_0^{-1}\mathbf{m}_0. \tag{3.184}$$

Then, the logarithm of the integrand with respect to β can be written as

$$-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N)$$
$$+ \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal \mathbf{S}_0^{-1}\mathbf{m}_0$$
$$- \ln\Gamma(a_0) + a_0\ln b_0 + (a_0 - 1)\ln\beta - b_0\beta \tag{3.185}$$
$$= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N)$$
$$- \ln\Gamma(a_0) + a_0\ln b_0 + (a_N - 1)\ln\beta - b_N\beta,$$

where

$$a_N = a_0 + \frac{N}{2},$$
$$b_N = b_0 + \frac{\beta}{2}\|\mathbf{t}\|^2 + \frac{\beta}{2}\mathbf{m}_0^\intercal \mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N. \tag{3.186}$$

Therefore, the logarithm of the integral with respect to β can be written as

$$-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N) - \ln\Gamma(a_0) + a_0\ln b_0 + \ln\Gamma(a_N) - a_N\ln b_N. \tag{3.187}$$

Thus,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left(\frac{\det \mathbf{S}_N}{\det \mathbf{S}_0} \right)^{\frac{1}{2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_N^{a_N}}. \tag{3.188}$$

3.24

Refer to 3.12.

4 Linear Models for Classification

4.1

Let $x_1, \cdots, x_M$ and $y_1, \cdots, y_N$ be two sets of data points. Then, the corresponding convex hulls are defined as the sets of all points $\mathbf{x}$ and $\mathbf{y}$ such that

$$
\begin{aligned}
\mathbf{x} &= \sum_{m=1}^{M} \alpha_m \mathbf{x}_m, \\
\mathbf{y} &= \sum_{n=1}^{N} \beta_n \mathbf{y}_n,
\end{aligned}
\tag{4.1}
$$

where

$$
\sum_{m=1}^{M} \alpha_m = \sum_{n=1}^{N} \beta_n = 1,
\tag{4.2}
$$

$$
\alpha_m \geq 0, \beta_n \geq 0.
$$

Let us assume that $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ below are subject to the constraints above.

If the convex hulls intersect, then there exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$
\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n.
\tag{4.3}
$$

Then,

$$
\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\mathsf{T} \mathbf{x}_m + w_0 \right) = \hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m} \alpha_m,
\tag{4.4}
$$

for any $\hat{\mathbf{w}}$ and w_0. The right hand side can be written as

$$
\hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n = \sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\mathsf{T} \mathbf{y}_n + w_0 \right).
\tag{4.5}
$$

Therefore, there do not exist $\hat{\mathbf{w}}$ and w_0 such that

$$
\begin{aligned}
\hat{\mathbf{w}}^\mathsf{T} \mathbf{x}_m + w_0 &> 0, \\
\hat{\mathbf{w}}^\mathsf{T} \mathbf{y}_n + w_0 &< 0.
\end{aligned}
\tag{4.6}
$$

Conversely, if there exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}
\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0 &> 0, \\
\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0 &< 0,
\end{aligned} \tag{4.7}$$

then

$$\begin{aligned}
\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0 \right) &> 0, \\
\sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0 \right) &< 0.
\end{aligned} \tag{4.8}$$

The left hand sides can be written as

$$\begin{aligned}
\hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m=1}^{M} \alpha_m &= \hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0, \\
\hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n &= \hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0.
\end{aligned} \tag{4.9}$$

Therefore, there do not exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n. \tag{4.10}$$

Thus, the convex hulls do not intersect.

4.2 (Incomplete)

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ and $\mathbf{w}_1, \cdots, \mathbf{w}_K$ are variables in M dimensions and $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are ones in K dimensions. Let

$$E(\tilde{\mathbf{W}}) = \frac{1}{2} \operatorname{tr} \left((\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^\mathsf{T} (\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}) \right), \tag{4.11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\mathsf{T} \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\mathsf{T} \end{bmatrix},$$

$$\tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\mathsf{T} \\ \vdots \\ \mathbf{t}_N^\mathsf{T} \end{bmatrix}.$$

Setting the derivative with respect to $\tilde{\mathbf{W}}$ to zero gives

$$\mathbf{O} = \tilde{\mathbf{X}}^\mathsf{T}(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}). \tag{4.12}$$

Therefore,

$$\operatorname*{argmin}_{\tilde{\mathbf{w}}} E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\mathsf{T}\mathbf{T}. \tag{4.13}$$

Let $\tilde{\mathbf{W}}^*$ denote the least-square solution above. If

$$\mathbf{a}^\mathsf{T}\mathbf{t}_n + b = 0, \tag{4.14}$$

for $n = 1, \cdots, N$, then

$$\mathbf{a}^\mathsf{T}(\tilde{\mathbf{W}}^*)^\mathsf{T}\tilde{\mathbf{x}} + b = 0? \tag{4.15}$$

where $\tilde{\mathbf{x}}$ is a variable in $M + 1$ dimensions whose first element is 1.

$$(\tilde{\mathbf{W}}^*)^\mathsf{T}\tilde{\mathbf{x}} = \mathbf{T}^\mathsf{T}\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}}. \tag{4.16}$$