Chapter 4

Multivariate Random Variables

In this chapter, we will review some topics related to random vectors, which will be of use in the following chapters.

4.1 Gaussian Random Vectors (Multivariate Normal Distribution)

Recall that a random variable X is Gaussian (normal) with mean μ and variance $\sigma^2 > 0$ if the pdf of X is given by

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}. \tag{4.1}$$

Definition 4.1. A collection of random variables is **jointly Gaussian** if any linear combination of these variables is Gaussian. A **Gaussian random vector**, also known as a *multivariate normal vector*, is a vector whose elements are jointly Gaussian. A collection of random vectors is jointly Gaussian if the vector obtained by concatenating them is jointly Gaussian.

Example 4.2. If $\begin{pmatrix} X \\ Y \end{pmatrix}$ is a Gaussian vector, then $Z = 2X + 3Y$ is Gaussian. Furthermore,

$$\mathbb{E}[Z] = 2\,\mathbb{E}[X] + 3\,\mathbb{E}[Y], \tag{4.2}$$

$$\operatorname{Var}(Z) = \operatorname{Cov}(2X+3Y, 2X+3Y) = 4\operatorname{Cov}(X,X) + 12\operatorname{Cov}(X,Y) + 9\operatorname{Cov}(Y,Y) \tag{4.3}$$

$$= 4\operatorname{Var}(X) + 12\operatorname{Cov}(X,Y) + 9\operatorname{Var}(Y), \tag{4.4}$$

which completely characterizes the distribution of Z as $Z \sim \mathcal{N}(\mathbb{E}[Z], \operatorname{Var}(Z))$. △

For a Gaussian random vector $\boldsymbol{X}$ of dimension d, with mean $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and covariance matrix $\mathsf{K} = \operatorname{Cov}(\boldsymbol{X}) = \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T]$, we have

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\mathsf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \mathsf{K}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{4.5}$$

provided that the covariance matrix is invertible.

The elements of $\boldsymbol{X}$ are **independent** if and only if the covariance matrix is diagonal.

4.1.1 Maximum likelihood estimation

Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with distribution $\mathcal{N}(\boldsymbol{\theta}^*, \mathsf{K}^*)$ given in (4.5), where $\boldsymbol{\theta}^*, \mathsf{K}^*$ are unknown. Suppose we are interested in the relationship between X_d and $X_1, \ldots, X_{d-1}$. For example, for $\boldsymbol{X}^T = (X_1, X_2, X_3)$, X_1 and X_2 could indicate the heights of the parents and X_3 could be the height of the child. We may, for example, be interested in finding $\mathbb{E}[X_d|X_1,\dots,X_{d-1}]$, thus estimating X_d based on $X_1,\dots,X_{d-1}$. If we find the distribution, in other words, $\boldsymbol{\theta}^*, \mathsf{K}^*$, we can do so. Furthermore, the matrix K^* can indicate which dimensions are more strongly correlated.

Consider a set of n iid samples $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$, where each $\boldsymbol{x}_i$ is a sample of $\boldsymbol{X}$. We denote the elements of $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})$.

To estimate $\boldsymbol{\theta}^*$ and K^*, we write

$$\ell(\boldsymbol{\theta}, \mathsf{K}) = \ln p(\mathcal{D}; \boldsymbol{\theta}, \mathsf{K}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i; \boldsymbol{\theta}, \mathsf{K}) \tag{4.6}$$

$$\doteq \frac{n}{2} \ln |\mathsf{K}^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})^T \mathsf{K}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta}), \tag{4.7}$$

where we have used the fact that $|\mathsf{K}^{-1}| = \frac{1}{|\mathsf{K}|}$.

As seen in the appendix (last chapter), for a symmetric matrix A, we have $\frac{d}{d\boldsymbol{v}}(\boldsymbol{y}^T \mathsf{A} \boldsymbol{y}) = 2\boldsymbol{y}^T \mathsf{A} \frac{d\boldsymbol{y}}{d\boldsymbol{v}}$. Hence,

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = -\frac{1}{2} \sum_{i=1}^{n} 2(\boldsymbol{x}_i - \boldsymbol{\theta})^T \mathsf{K}^{-1}(-\mathsf{I}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})^T \mathsf{K}^{-1}. \tag{4.8}$$

Setting this equal to zero yields

$$\hat{\boldsymbol{\theta}}_{ML} = \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i. \tag{4.9}$$

Exercise 4.3. Using the facts

$$\frac{\partial}{\partial \mathsf{A}} \boldsymbol{x}^T \mathsf{A} \boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}^T, \quad \frac{\partial}{\partial \mathsf{A}} \ln |\mathsf{A}| = \mathsf{A}^{-T} \tag{4.10}$$

prove that

$$\hat{\mathsf{K}}_{ML} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \tag{4.11}$$

△

4.1.2 Bayesian estimation

We now solve the same problem using Bayesian estimation, with the following likelihood

$$\boldsymbol{X}|\boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\Theta}, \mathsf{K}), \tag{4.12}$$

$$p(\boldsymbol{x}_1^n|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})^T \mathsf{K}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})\right), \tag{4.13}$$

where, for simplicity, we assume K is known and we only need to estimate $\boldsymbol{\Theta}$. As the prior, we choose

$$\boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathsf{S}_0) \tag{4.14}$$

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \mathsf{S}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right). \tag{4.15}$$

Hence,

$$p(\boldsymbol{\theta}|\boldsymbol{x}_1^n) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \mathsf{S}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0) - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})^T \mathsf{K}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})\right). \tag{4.16}$$

The exponent in the posterior is quadratic in $\boldsymbol{\theta}$, indicating that $\boldsymbol{\Theta}$ has a Gaussian distribution. So $\boldsymbol{\Theta}|\boldsymbol{x}_1^n \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_n, \mathsf{S}_n)$, for appropriate choices of $\hat{\boldsymbol{\theta}}_n$ and S_n,

$$p(\boldsymbol{\theta}|\boldsymbol{x}_1^n) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)^T \mathsf{S}_n^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\right). \tag{4.17}$$

To find $\hat{\boldsymbol{\theta}}_n$ and S_n, we equate (4.16) and (4.17), ignoring constant multiplicative factors, which leads to

$$(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T \mathsf{S}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0) + \sum_{i=1}^n (\boldsymbol{\theta}-\boldsymbol{x}_i)^T \mathsf{K}^{-1}(\boldsymbol{\theta}-\boldsymbol{x}_i) \doteq (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)^T \mathsf{S}_n^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n), \tag{4.18}$$

$$\boldsymbol{\theta}^T \mathsf{S}_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathsf{S}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\theta}^T \mathsf{K}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathsf{K}^{-1}\sum_{i=1}^n \boldsymbol{x}_i \doteq \boldsymbol{\theta}^T \mathsf{S}_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathsf{S}_n^{-1}\hat{\boldsymbol{\theta}}_n. \tag{4.19}$$

Here, we have used the fact that

$$(\boldsymbol{a}-\boldsymbol{b})^T \mathsf{A}(\boldsymbol{a}-\boldsymbol{b}) = \boldsymbol{a}^T\mathsf{A}\boldsymbol{a} - \boldsymbol{a}^T\mathsf{A}\boldsymbol{b} - \boldsymbol{b}^T\mathsf{A}\boldsymbol{a} + \boldsymbol{b}^T\mathsf{A}\boldsymbol{b} = \boldsymbol{a}^T\mathsf{A}\boldsymbol{a} - 2\boldsymbol{a}^T\mathsf{A}\boldsymbol{b} + \boldsymbol{b}^T\mathsf{A}\boldsymbol{b},$$

for vectors $\boldsymbol{a}, \boldsymbol{b}$ and a symmetric matrix A. Note that $\boldsymbol{a}^T\mathsf{A}\boldsymbol{b} = \boldsymbol{b}^T\mathsf{A}\boldsymbol{a}$, as both sides are scalars and $\boldsymbol{a}^T\mathsf{A}\boldsymbol{b} = (\boldsymbol{a}^T\mathsf{A}\boldsymbol{b})^T = \boldsymbol{b}^T\mathsf{A}\boldsymbol{a}$.

We now collect the terms of the form $\boldsymbol{\theta}^T\mathsf{A}\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^T(\mathsf{S}_0^{-1} + n\mathsf{K}^{-1})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T(\mathsf{S}_0^{-1}\boldsymbol{\mu}_0 + \mathsf{K}^{-1}\sum_{i=1}^n \boldsymbol{x}_i) \doteq \boldsymbol{\theta}^T \mathsf{S}_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T \mathsf{S}_n^{-1}\hat{\boldsymbol{\theta}}_n, \tag{4.20}$$

leading to the following values for the parameters of the posterior distribution $\boldsymbol{\Theta}|\boldsymbol{x}_1^n \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_n, \mathsf{S}_n)$,

$$\mathsf{S}_n^{-1} = \mathsf{S}_0^{-1} + n\mathsf{K}^{-1}, \tag{4.21}$$

$$\hat{\boldsymbol{\theta}}_n = \mathsf{S}_n(\mathsf{S}_0^{-1}\boldsymbol{\mu}_0 + n\mathsf{K}^{-1}\bar{\boldsymbol{x}}) \tag{4.22}$$

$$= (\mathsf{S}_0^{-1} + n\mathsf{K}^{-1})^{-1}(\mathsf{S}_0^{-1}\boldsymbol{\mu}_0 + n\mathsf{K}^{-1}\bar{\boldsymbol{x}}), \tag{4.23}$$

where $\bar{\boldsymbol{x}}$ is $\sum_{i=1}^n \boldsymbol{x}_i/n$. The posterior mean, $\hat{\boldsymbol{\theta}}_n$, which we can also view as a point estimate, is the weighted average of the prior mean $\boldsymbol{\mu}_0$ and what is suggested by the data $\bar{\boldsymbol{x}}$.

Exercise 4.4. Find $\hat{\boldsymbol{\theta}}_n$ and S_n^{-1} when $\mathsf{S}_0 = s^2\mathsf{I}$ and $\mathsf{K} = \sigma^2\mathsf{I}$ and interpret the results. △