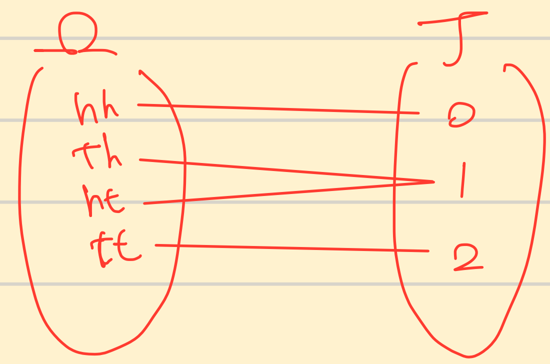
6.1 Construction of a Probability Space

- Random variable : function.

  $X : \Omega \to T$

  Mapping from Ω to T



$X(th) = 1 \iff X = 1$

* Ω : sample space (= state space)

: set of all possible outcomes of experiment.

* T : target space

: Set of quantity of interest.

6.2 Discrete and Continuous Probabilities

- Probability density function : $f : \mathbb{R}^D \to \mathbb{R}$ is called probability density function if

$$\left[\begin{array}{l} x \in \mathbb{R}^D, \ \forall x \ f(x) \geq 0 \\ \int f(x) dx = 1 \end{array}\right.$$

* if random variable x is discrete, we can replace integral with sum.

6.3 Sum Rule, Product Rule, and Bayes' Theorem

- Sum Rule

$$p(x) = \sum_{i=1}^{n} p(x, y_i)$$

$$= \int_y p(x, y)\, dy$$

$$p(x_i) = \int p(x_1, x_2, \ldots, x_n)\, dx_{\setminus i}$$ → $\forall x$ except i.

- produce rule

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

- Bayes rule

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(y|x) = \frac{p(y|x)\,p(x)}{p(x)} = \frac{p(x|y)\,p(y)}{p(x)}$$

* $p(y|x)$ 와 $p(y)$의 차이.

$p(y|x)$ is something we know after having observed 'x'.

6.4 Summary Statistics and Independence

Example 6.4

Consider the two-dimensional distribution illustrated in Figure 6.4:

$$p(x) = 0.4\,\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \underbrace{\begin{bmatrix}10\\2\end{bmatrix}}_{\mu}, \underbrace{\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}}_{\Sigma}\right) + 0.6\,\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \underbrace{\begin{bmatrix}0\\0\end{bmatrix}}_{\mu}, \underbrace{\begin{bmatrix}8.4 & 2.0\\2.0 & 1.7\end{bmatrix}}_{\Sigma}\right). \tag{6.33}$$

We will define the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ in Section 6.5. Also shown is its corresponding marginal distribution in each dimension. Observe that the distribution is bimodal (has two modes), but one of the marginal distributions is unimodal (has one mode). The horizontal bimodal univariate distribution illustrates that the mean and median can be different from each other. While it is tempting to define the two-dimensional median to be the concatenation of the medians in each dimension, the fact that we cannot define an ordering of two-dimensional points makes it difficult. When we say "cannot define an ordering", we mean that there is more than one way to define the relation $<$ so that $\begin{bmatrix}3\\0\end{bmatrix} < \begin{bmatrix}2\\3\end{bmatrix}$.

Multi-modal이면,
mean, median, modes
간 관계 일률적 X

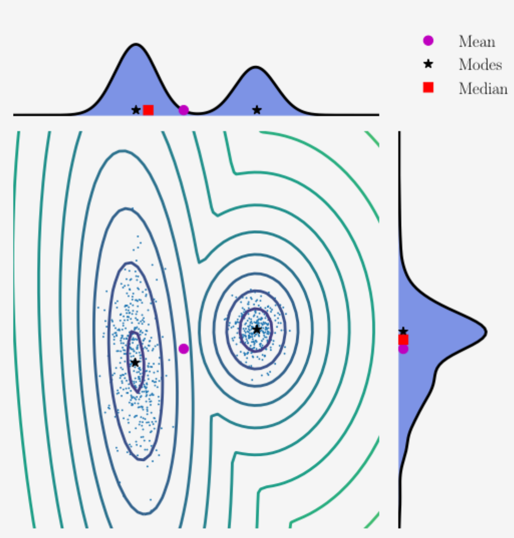


Figure 6.4 Illustration of the mean, mode, and median for a two-dimensional dataset, as well as its marginal densities.

$$\mathrm{Cov}_{X,Y}[x,y] := \mathbb{E}_{X,Y}\big[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])\big]\,.$$

$$\mathrm{Cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]\,.$$

univariate

$$\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^\top = \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}\,.$$

$\Leftarrow$ multivariate

$$\mathbb{V}_X[\boldsymbol{x}] = \mathrm{Cov}_X[\boldsymbol{x},\boldsymbol{x}] \tag{6.38a}$$

$$= \mathbb{E}_X[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top \tag{6.38b}$$

$$= \begin{bmatrix} \mathrm{Cov}[x_1,x_1] & \mathrm{Cov}[x_1,x_2] & \dots & \mathrm{Cov}[x_1,x_D] \\ \mathrm{Cov}[x_2,x_1] & \mathrm{Cov}[x_2,x_2] & \dots & \mathrm{Cov}[x_2,x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D,x_1] & \dots & \dots & \mathrm{Cov}[x_D,x_D] \end{bmatrix}. \tag{6.38c}$$

Covariance matrix
$\Rightarrow$ symmetric
positive-semidefinite

Correlation
= normalized version of covariance $= \rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

$$\mathbb{E}_Y[\boldsymbol{y}] = \mathbb{E}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\,, \tag{6.50}$$

$$\mathbb{V}_Y[\boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] = \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_X[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\,, \tag{6.51}$$

respectively. Furthermore,

$$\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}]^\top \tag{6.52a}$$

$$= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top \tag{6.52b}$$

$$= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top \tag{6.52c}$$

$$\overset{(6.38b)}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top\,, \tag{6.52d}$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ is the covariance of X.

Definition 6.11 (Conditional Independence)**.** Two random variables X and Y are *conditionally independent* given Z if and only if

$$p(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{z}) = p(\boldsymbol{x}\,|\,\boldsymbol{z})p(\boldsymbol{y}\,|\,\boldsymbol{z}) \quad \text{for all} \quad \boldsymbol{z} \in \mathcal{Z}\,, \tag{6.55}$$

6.5 Gaussian Distribution

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

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (6.63)$$

Marginal, conditional of $\mathcal{N}$ $\approx$ Gaussian distribution

Produce, Sum, Linear transformation of $\mathcal{N}$

$\approx$ Gaussian distribution

$$p(\boldsymbol{x}+\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y,\ \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right).$$

Theorem 6.12. *Consider a mixture of two univariate Gaussian densities*

$$p(x) = \alpha p_1(x) + (1-\alpha) p_2(x), \quad (6.80)$$

where the scalar $0 < \alpha < 1$ is the mixture weight, and $p_1(x)$ and $p_2(x)$ are univariate Gaussian densities (Equation (6.62)) with different parameters, i.e., $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$.

Then the mean of the mixture density $p(x)$ is given by the weighted sum of the means of each random variable:

$$\mathbb{E}[x] = \alpha\mu_1 + (1-\alpha)\mu_2. \quad (6.81)$$

The variance of the mixture density $p(x)$ is given by

$$\mathbb{V}[x] = \left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right] + \left(\left[\alpha\mu_1^2 + (1-\alpha)\mu_2^2\right] - \left[\alpha\mu_1 + (1-\alpha)\mu_2\right]^2\right). \quad (6.82)$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}. \quad (6.86)$$

Similarly the variance of $\boldsymbol{y}$ can be found by using (6.51):

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top. \quad (6.87)$$

This means that the random variable $\boldsymbol{y}$ is distributed according to

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu},\ \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right). \quad (6.88)$$

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \iff (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y} = \boldsymbol{x}. \quad (6.90)$$

Hence, $\boldsymbol{x}$ is a linear transformation of $\boldsymbol{y}$, and we obtain

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y},\ (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\right). \quad (6.91)$$