مشخصه‌های عددی

Numeric Attributes

$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Empirical Cumulative Distribution Function

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \qquad I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

Inverse Cumulative Distribution Function

$$F^{-1}(q) = \min\{x \mid F(x) \geq q\} \qquad \text{for } q \in [0,1]$$

Empirical Probability Mass Function

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x) \qquad I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

Measures of Central Tendency

Mean

$$\boxed{\mu = E[X] = \sum_{x} x \cdot f(x)}$$

$$\boxed{\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx}$$

Sample Mean

$$\boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$$\hat{\mu} = \sum_{x} x \cdot \hat{f}(x) = \sum_{x} x \left(\frac{1}{n} \sum_{i=1}^{n} I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample Mean Is Unbiased

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

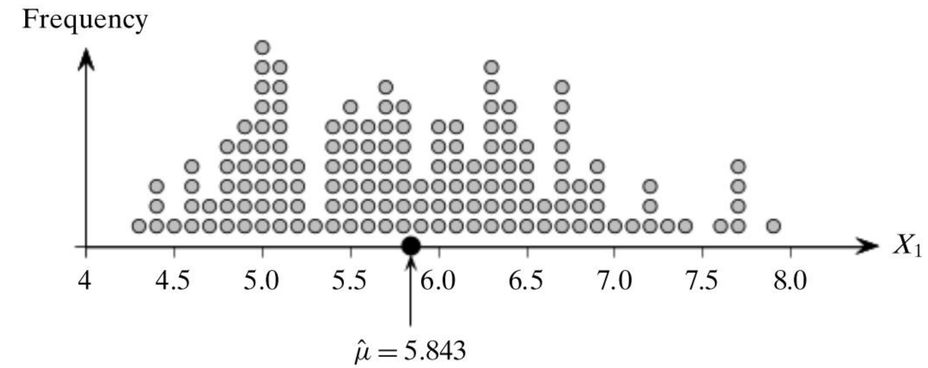
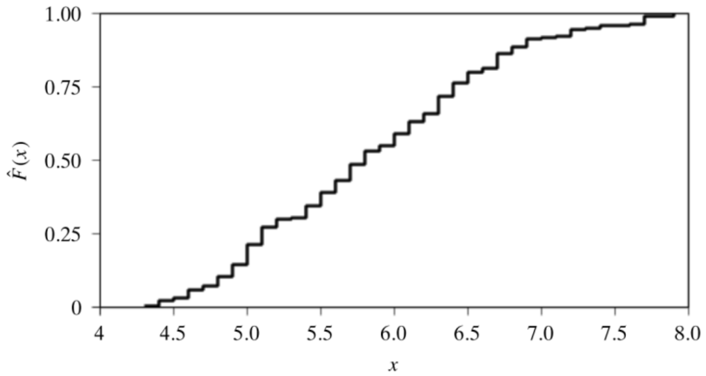
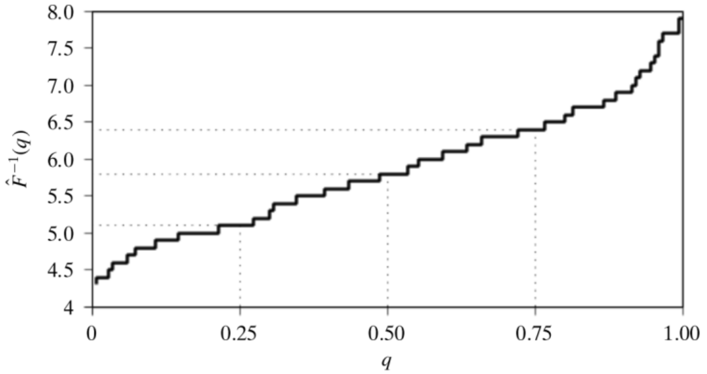


Figure 2.2. Sample mean for `sepal length`. Multiple occurrences of the same value are shown stacked.



(a) Empirical CDF



(b) Empirical inverse CDF

Measures of Dispersion

Range $$r = \max\{X\} - \min\{X\}$$

The sample range is a statistic, given as $$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}$$

Interquartile Range

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25)$$

The *sample IQR*

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

Variance and Standard Deviation

$$\boxed{\sigma^2 = \text{var}(X) = E[(X-\mu)^2]} = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

$$\begin{aligned}\sigma^2 = \text{var}(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2\end{aligned}$$

Sample Variance

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

The standard score, also called the z-score

$$\boxed{z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}}$$

Variance of the Sample Mean

$$\text{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n}\text{var}(x_i) = \sum_{i=1}^{n}\sigma^2 = n\sigma^2 \qquad E\left[\sum_{i=1}^{n} x_i\right] = n\mu$$

$$\begin{aligned}\text{var}(\hat{\mu}) = E[(\hat{\mu}-\mu)^2] = E[\hat{\mu}^2] - \mu^2 &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \frac{1}{n^2}E\left[\sum_{i=1}^{n} x_i\right]^2 \\ &= \frac{1}{n^2}\left(E\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] - E\left[\sum_{i=1}^{n} x_i\right]^2\right) = \frac{1}{n^2}\,\text{var}\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Bias of Sample Variance

$$\sum_{i=1}^{n}(x_i-\mu)^2 = n(\hat{\mu}-\mu)^2 + \sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right] - E[(\hat{\mu}-\mu)^2]$$

Recall that the random variables xi are IID according to X, which means that they have the same mean μ and variance σ^2 as X.

$$E[(x_i-\mu)^2] = \sigma^2 \quad E[\hat{\sigma}^2] = \frac{1}{n}n\sigma^2 - \frac{\sigma^2}{n}$$

$$= \left(\frac{n-1}{n}\right)\sigma^2$$

It is asymptotically unbiased

$$E[\hat{\sigma}^2] \to \sigma^2 \qquad \text{as } n \to \infty$$

An unbiased estimate of the sample variance

$$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

$$E[\hat{\sigma}_u^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] = \frac{1}{n-1}\cdot E\left[\sum_{i=1}^{n}(x_i-\mu)^2\right] - \frac{n}{n-1}\cdot E[(\hat{\mu}-\mu)^2]$$

$$= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\cdot\frac{\sigma^2}{n}$$

$$= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2$$

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Empirical Joint Probability Mass Function

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^{n} I(x_{i1} = x_1, x_{i2} = x_2)$$

Measures of Location and Dispersion

Mean $\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left(\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

Variance $\text{var}(\mathbf{D}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2$

Measures of Association

Covariance

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$

The sample covariance

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

Correlation

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

The sample correlation for attributes X1 and X2

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$$

Geometric Interpretation of Sample Covariance and Correlation

$$\hat{\sigma}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{n}$$

$$\hat{\rho}_{12} = \frac{\bar{X}_1^T \bar{X}_2}{\sqrt{\bar{X}_1^T \bar{X}_1} \sqrt{\bar{X}_2^T \bar{X}_2}} = \frac{\bar{X}_1^T \bar{X}_2}{\|\bar{X}_1\| \, \|\bar{X}_2\|} = \left(\frac{\bar{X}_1}{\|\bar{X}_1\|} \right)^T \left(\frac{\bar{X}_2}{\|\bar{X}_2\|} \right) = \cos\theta$$

Covariance Matrix

$$
\begin{aligned}
\boldsymbol{\Sigma} &= E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] \\
&= E\left[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}\begin{pmatrix} X_1-\mu_1 & X_2-\mu_2 \end{pmatrix}\right] \\
&= \begin{pmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}
\end{aligned}
$$

Because $\sigma_{12} = \sigma_{21}$, $\boldsymbol{\Sigma}$ is a *symmetric* matrix.

The *total variance* of the two attributes $tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2$

$$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 = (1-\rho_{12}^2)\sigma_1^2\sigma_2^2$$

The sample covariance matrix

$$\boxed{\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}}$$

$$\boxed{\text{var}(\mathbf{D}) = tr(\widehat{\boldsymbol{\Sigma}}) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix} = \begin{pmatrix} — & \mathbf{x}_1^T & — \\ — & \mathbf{x}_2^T & — \\ & \vdots & \\ — & \mathbf{x}_n^T & — \end{pmatrix}$$

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \in \mathbb{R}^d \qquad X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in \mathbb{R}^n$$

Mean

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

$$\boxed{\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i}$$

Covariance Matrix

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Covariance Matrix Is Positive Semidefinite

$\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \geq 0$ for any d-dimensional vector $\mathbf{a}$

$$\begin{aligned} \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} &= \mathbf{a}^T E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \mathbf{a} \\ &= E[\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{a}] \\ &= E[Y^2] \\ &\geq 0 \end{aligned}$$

Because Σ is also symmetric, this implies that all the eigenvalues of Σ are real and non-negative. In other words the d eigenvalues of Σ can be arranged from the largest to the smallest as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$$

Total and Generalized Variance

$$tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2$$

$$\det(\mathbf{\Sigma}) = |\mathbf{\Sigma}| = \prod_{i=1}^{d} \lambda_i$$

Since all the eigenvalues of $\mathbf{\Sigma}$ are non-negative ($\lambda_i \geq 0$), it follows that $\det(\mathbf{\Sigma}) \geq 0$.

Sample Covariance Matrix

$$\widehat{\mathbf{\Sigma}} = E[(\mathbf{X} - \hat{\boldsymbol{\mu}})(\mathbf{X} - \hat{\boldsymbol{\mu}})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \hat{} & \sigma_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \hat{} & \sigma_{2d} \\ \cdots & \cdots \cdots & \cdots & \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots \hat{} & \sigma_d^2 \end{pmatrix}$$

The sample covariance matrix is thus given as the pairwise inner or dot products of the centered attribute vectors, normalized by the sample size.

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\left(\overline{\mathbf{D}}^T \, \overline{\mathbf{D}}\right) = \frac{1}{n} \begin{pmatrix} \overline{X}_1^T \overline{X}_1 & \overline{X}_1^T \overline{X}_2 & \cdots & \overline{X}_1^T \overline{X}_d \\ \overline{X}_2^T \overline{X}_1 & \overline{X}_2^T \overline{X}_2 & \cdots & \overline{X}_2^T \overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T \overline{X}_1 & \overline{X}_d^T \overline{X}_2 & \cdots & \overline{X}_d^T \overline{X}_d \end{pmatrix}$$

In terms of the centered points x_i, the sample covariance matrix can also be written as a sum of rank-one matrices obtained as the outer product of each centered point:

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \overline{\mathbf{x}}_i \cdot \overline{\mathbf{x}}_i^T$$

Sample Scatter Matrix

$$\mathbf{S} = \overline{\mathbf{D}}^T \, \overline{\mathbf{D}} = \sum_{i=1}^{n} \overline{\mathbf{x}}_i \cdot \overline{\mathbf{x}}_i^T \qquad \mathbf{S} = n \cdot \widehat{\mathbf{\Sigma}}$$

Range Normalization

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

Standard Score Normalization

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

$$\hat{\mu}' = 0$$
$$\hat{\sigma}' = 1$$

Table 2.1. Dataset for normalization

$\mathbf{x}_i$	Age (X_1)	Income (X_2)
$\mathbf{x}_1$	12	300
$\mathbf{x}_2$	14	500
$\mathbf{x}_3$	18	1000
$\mathbf{x}_4$	23	2000
$\mathbf{x}_5$	27	3500
$\mathbf{x}_6$	28	4000
$\mathbf{x}_7$	34	4300
$\mathbf{x}_8$	37	6000
$\mathbf{x}_9$	39	2500
$\mathbf{x}_{10}$	40	2700

$$\texttt{Age}' = (0, 0.071, 0.214, 0.393, 0.536, 0.571, 0.786, 0.893, 0.964, 1)^T$$

$$\texttt{Income}' = (0, 0.035, 0.123, 0.298, 0.561, 0.649, 0.702, 1, 0.386, 0.421)^T$$

	Age	Income
$\hat{\mu}$	27.2	2680
$\hat{\sigma}$	9.77	1726.15

$$\texttt{Age}' = (-1.56, -1.35, -0.94, -0.43, -0.02, 0.08, 0.70, 1.0, 1.21, 1.31)^T$$

$$\texttt{Income}' = (-1.38, -1.26, -0.97, -0.39, 0.48, 0.77, 0.94, 1.92, -0.10, 0.01)^T$$

Univariate Normal Distribution

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

Probability Mass

$$P(a \le x \le b) = \int_a^b f(x|\mu,\sigma^2)\,dx$$

we are often interested in the probability mass concentrated within k standard deviations from the mean

$$z = \frac{x-\mu}{\sigma}$$

$$P(-k \le z \le k) = \frac{1}{\sqrt{2\pi}} \int_{-k}^{k} e^{-\frac{1}{2}z^2} dz = \frac{2}{\sqrt{2\pi}} \int_0^k e^{-\frac{1}{2}z^2} dz$$

Multivariate Normal Distribution

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right\}$$

Geometry of the Multivariate Normal

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i \qquad \mathbf{u}_i \in \mathbb{R}^d$$

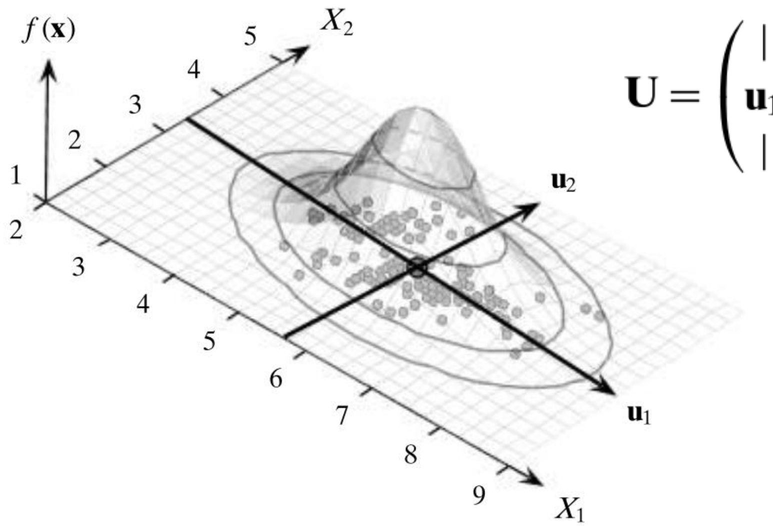
Because Σ is symmetric and positive semidefinite it has d real and non-negative eigenvalues, which can be arranged in order from the largest to the smallest as follows:

$$\lambda_1 \ge \lambda_2 \ge \cdots \lambda_d \ge 0$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

$$\mathbf{u}_i^T\mathbf{u}_i = 1 \quad \text{for all } i$$

$$\mathbf{u}_i^T\mathbf{u}_j = 0 \quad \text{for all } i \ne j$$



$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

Figure 2.8. Iris: `sepal length` and `sepal width`, bivariate normal density and contours.