Lecture 7 - Jan 31

Crash Course in Probability + Statistics

Reading

Data Mining and Machine Learning
2: Numerical Attributes

Upcoming Deadlines

Homework 2 (Feb 10)

A random experiment results in one of a number of outcomes from a sample space Ω

Subsets of the sample space $E \subseteq \Omega$ are events

A probability measure is a function $\mathbb{P}: \{\text{events}\} \to [0,1]$ where

1. $\mathbb{P}(\Omega) = 1$
2. If $A_1, A_2, \ldots$ are disjoint events,

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Properties of probability measures

$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ — A^c: A does not occur — probability of complement

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ — principle of

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$ Principle of inclusion-exclusion

Conditional Probability

The probability event A occurs given event B has occurred is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$ Conditional probability

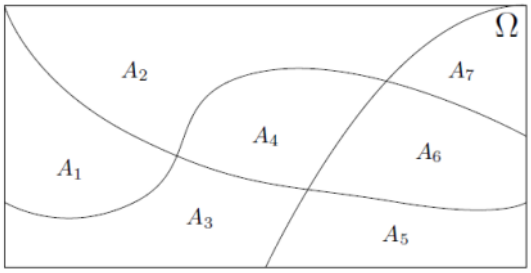
$\mathbb{P}(A)$ - prior probability $\mathbb{P}(A|B)$ - posterior probability

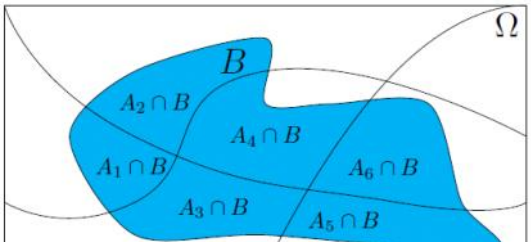
The Law of Total Probability

If events $A_1, A_2, \ldots$ partition the sample space Ω, then

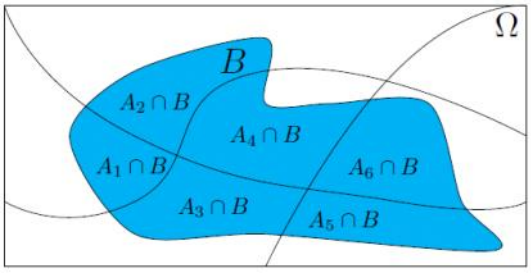
$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)$$

$$= \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$





$$= \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$



Bayes' Theorem

If $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Version 2 If $A_1, A_2, \ldots$ partition Ω,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}$$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Events A and B are ~~independent~~ if $P(A|B) = P(A)$

↳ this implies $P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$, i.e event B's occurrence has no impact on A.

A random variable is a measurable function $X: \Omega \to \mathbb{R}$ mapping outcomes to numbers.

Random variables valued in countable sets are discrete

Random variables valued in intervals (e.g. $\mathbb{R}$) are continuous

A probability distribution describes how likely a R.V. or set of RVs is to take different values.

In the discrete case, a probability mass function (PMF) f fully describes a probability distribution

$$f(x) = \mathbb{P}(X = x)$$

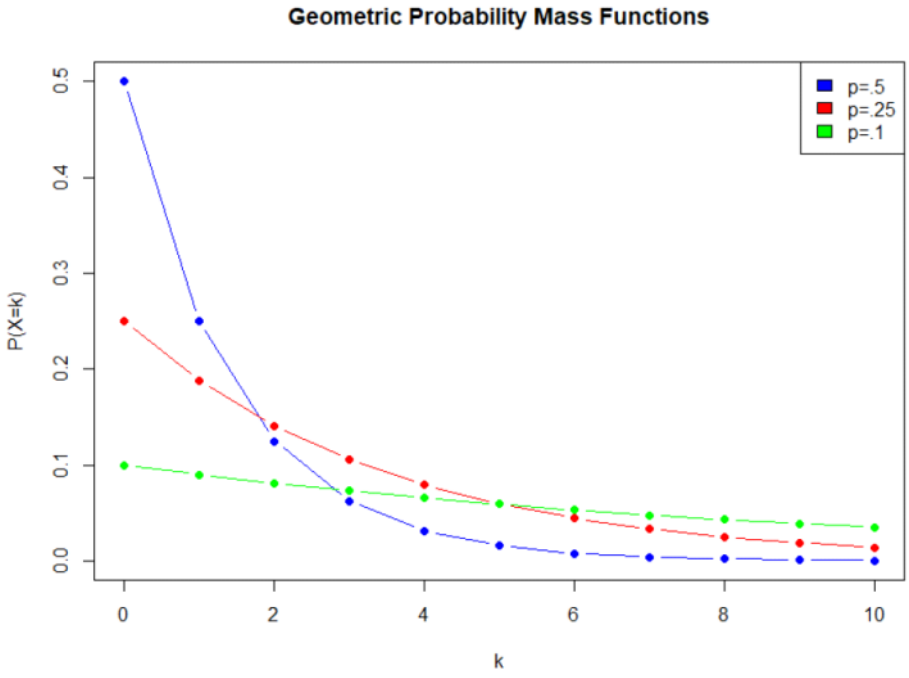
for each x in the range of X.

Example: Geometric R.V.

Let X be the number of failures before the first success in a sequence of independent binary experiments, each with

a sequence of independent binary experiments, each with probability p of success.

$$\mathbb{P}(X=k) = f(k) = \begin{cases} (1-p)^k p, & \text{if } k=0,1,2,\dots \\ 0, & \text{else} \end{cases}$$



In the continuous case, a probability density function (pdf) f fully describes a probability distribution

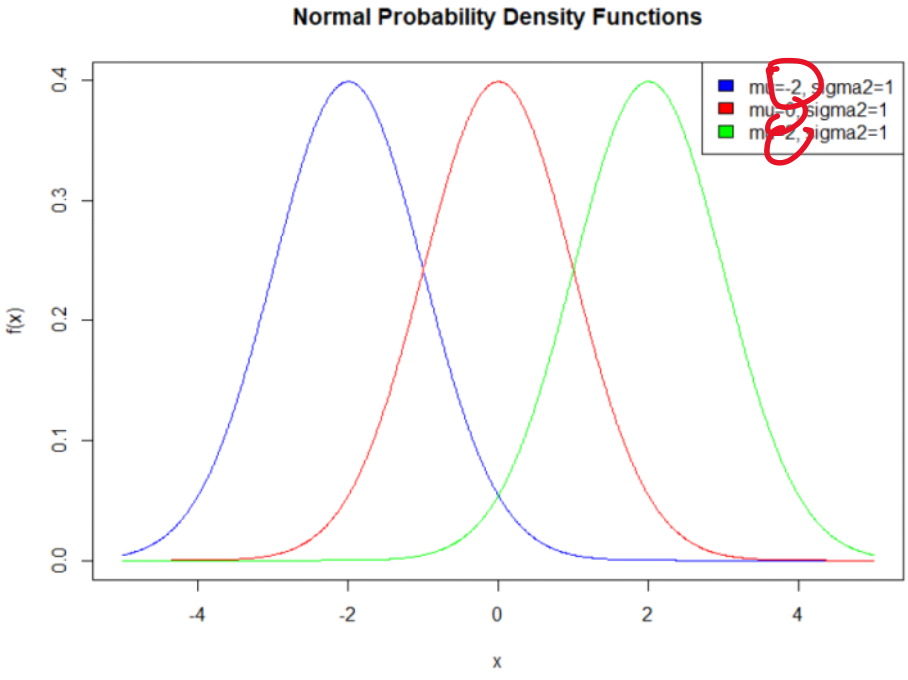
$$\int_E f(x)\,dx = \mathbb{P}(X \in E)$$

for a set $E \subseteq \mathbb{R}$.

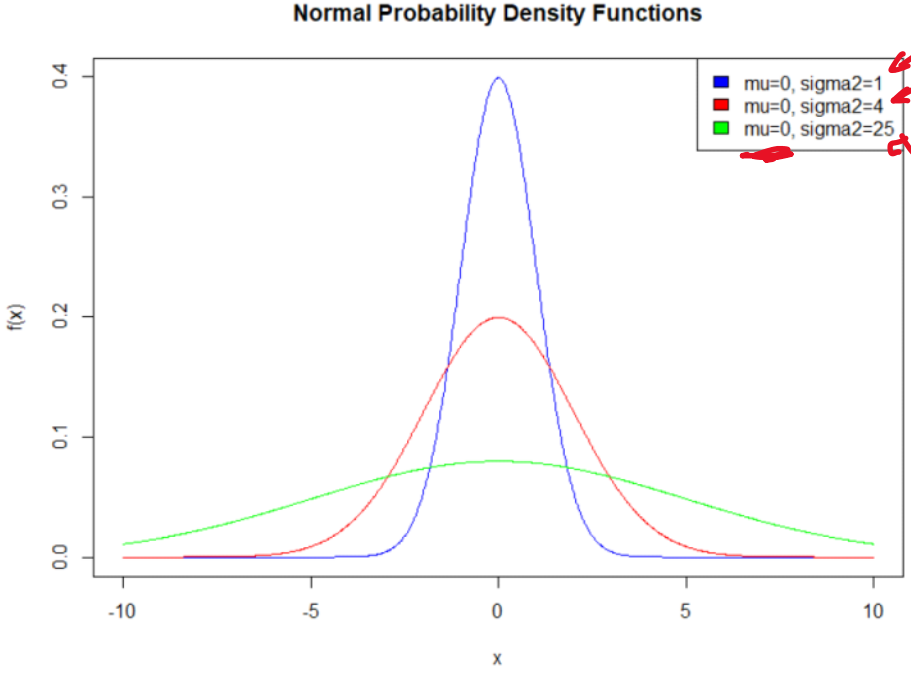
Example: Normal (Gaussian) RV

X has a normal distribution with parameters $\mu \in \mathbb{R}$, $\sigma > 0$ if its pdf is

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



μ is the center of the distribution



σ controls how dispersed the probability is

large σ → more "spread out"

small σ → more localized around μ

If $\mu=0, \sigma=1$, X is a standard normal R.V.

The expected value of a R.V. X with range R is

$$E[X] = \sum_{x \in R} x P(X=x) = \sum_{x \in R} x f(x) \quad \text{(discrete)}$$

$$E[X] = \int_R x f(x)\,dx \quad \text{(continuous)}$$

This is the mean value X takes

Properties of expectation

① If $g(X)$ is a R.V., $E[g(X)] = \begin{cases} \sum_{x \in R} g(x) f(x), & X \text{ is discrete} \\ \int_R g(x) f(x)\,dx, & X \text{ is continuous} \end{cases}$

② $E[aX+bY] = aE[X] + bE[Y]$

③ If X, Y are independent, $E[XY] = E[X]E[Y]$

The variance of X is $\mathrm{Var}(X) = E[(X - E[X])^2]$

SSE between X & its mean

↳ measures how "spread out" the pdf of X is

properties of variance

① $\mathrm{Var}(X) = E[X^2] - E[X]^2$

② $\mathrm{Var}(cX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(X)$ if X, Y are independent

If $X_1, X_2, \dots, X_n$ are discrete R.Vs, their **joint PMF** is

$$f(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

If they are continuous, the **joint pdf** is f such that

$$\int_E f(x_1, \dots, x_n)\, dx_1 \cdots dx_n = \mathbb{P}((X_1, \dots, X_n) \in E)$$

for any $E \subseteq \mathbb{R}^n$.

The expected value of some R.V. $g(X_1, \dots, X_n)$ is

$$E[g(X_1, \dots, X_n)] = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, \dots, x_n) f(x_1, \dots, x_n), & \text{if discrete} \\ \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, \dots, x_n) f(x_1, \dots, x_n)\, dx_1 \cdots dx_n, & \text{if continuous} \end{cases}$$

The **covariance** between X_j + X_k is

$$Cov(X_j, X_k) = E\left[(X_j - E[X_j])(X_k - E[X_k])\right] = \sigma_{jk}$$

Covariance quantifies how dependent $X_j + X_k$ are, but its value is dependent on the scale of $X_j + X_k$..

Correlation is a normalized version valued in $[-1,1]$.

The Covariance matrix for RVs $X_1, \dots, X_d$ is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

← square, symmetric matrix

The correlation of $X_j + X_k$ is

$$Corr(X_j, X_k) = \frac{Cov(X_j, X_k)}{\sigma_j \sigma_k}$$

Multivariate Normal (Gaussian) Distribution

A random vector $X = (X_1, \dots, X_d)$ has a MV. normal distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix Σ if its

Mahalanobis distance from x to μ ← takes into account variance & covar. info

with mean vector $\mu \in \mathbb{R}^k$ and

joint pdf is

Mahalanobis distance from x to μ taking into account variance & covar. info.

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

(x: vector)

X has a standard MV normal distribution if $\mu = \begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}$ and

$$\Sigma = I_d = \begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}$$

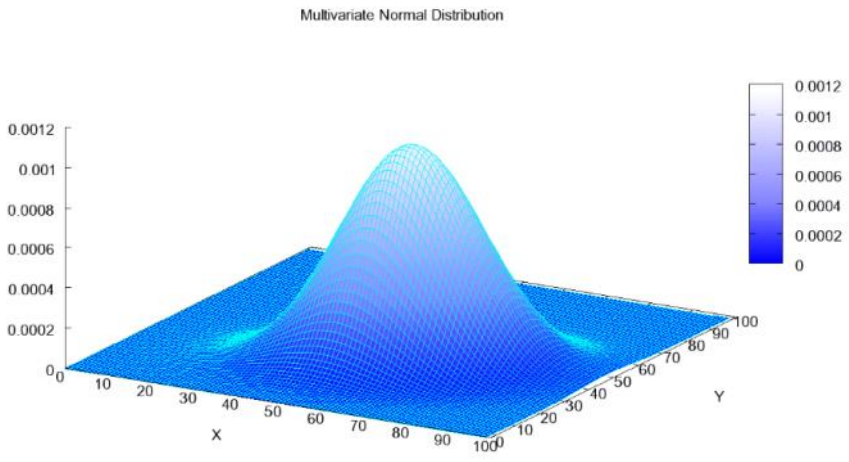
→ here, each X_i is an independent standard normal RV

$$\hookrightarrow f(x|0,I) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{(x-\mu)^T(x-\mu)}{2}\right) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{\|x-\mu\|^2}{2}\right)$$

If d=2, the multivariate normal distribution simplifies to

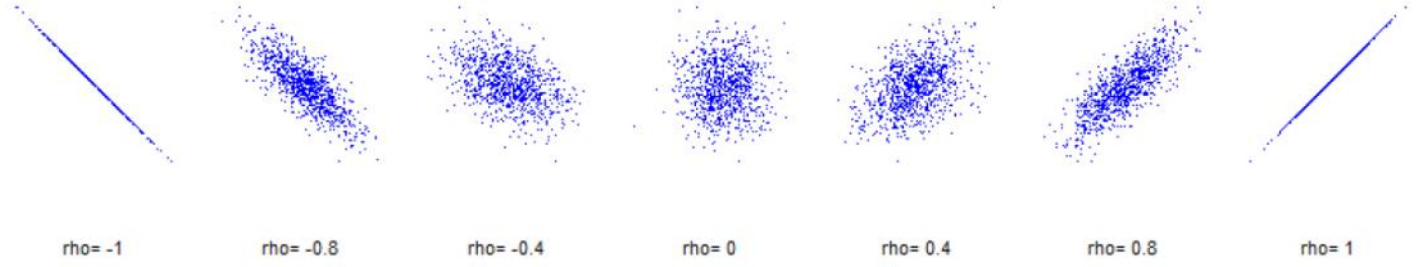
$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]}$$

If $\rho = 0$,



[BivariateNormal]

If ρ varies



A random sample is RVs $X_1, \dots, X_n$ that are independent and identically distributed (i.i.d.)

↳ a sample comes from some (usually unknown) population with some fixed but usually unknown distribution. Parameters of this distribution are population parameters

A sample statistic is a R.V. $g(\underbrace{X_1, \dots, X_n}_{\text{sample}})$ for some given g.

Univariate statistics

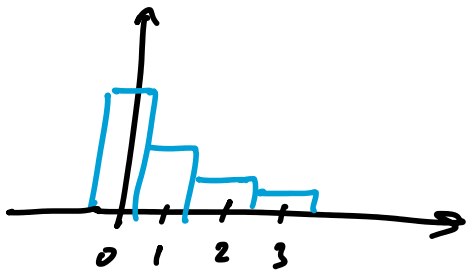
Let $D = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ be values of the RVs in a sample.

We treat it as a point in $\mathbb{R}^n$.

Let X be a R.V. with the population's distribution.

The empirical PMF of X is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i = x\}}$$



$\mathbb{1}_A = \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{else} \end{cases}$ is an indicator function

The sample mean of X is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

The centered feature vector of X is $\bar{X} = \begin{pmatrix} x_1 - \hat{\mu} \\ \vdots \\ x_n - \hat{\mu} \end{pmatrix}$

The sample mode of X is $\text{mode}(X) = \arg\max_x \hat{f}(x)$

The sample variance of X is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{1}{n}\|\bar{X}\|^2$

↳ note $\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\bar{X}^T\bar{X} = \frac{1}{n}\|\bar{X}\|^2$

The standard score (z-score) of x_i is $z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$

Multivariate Statistics

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \begin{matrix} \leftarrow x_1^T \\ \\ \leftarrow x_n^T \end{matrix}$$

$\uparrow$ X_1 $\uparrow$ X_2 $\uparrow$ X_d

Below, we give multivariate versions of sample statistics

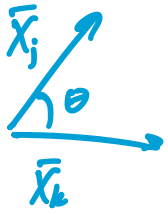
The sample mean vector is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_n \end{pmatrix}$ (x_i: vector)

The sample covariance between X_j & X_k is $\hat{\sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\hat{\mu}_j)(x_{ik}-\hat{\mu}_k) = \frac{1}{n}\bar{X}_j^T\bar{X}_k$

↳ notice $\sum(x_{ij}-\hat{\mu}_j)(x_{ik}-\hat{\mu}_k) = \bar{X}_j^T\bar{X}_k$

The sample correlation between X_j & X_k is $\frac{\hat{\sigma}_{jk}}{\hat{\sigma}_j\hat{\sigma}_k} = \frac{\bar{X}_j^T\bar{X}_k}{\|\bar{X}_j\|\cdot\|\bar{X}_k\|}$ ✱

$$\frac{\hat{\sigma}_{jk}}{\hat{\sigma}_j\hat{\sigma}_k} = \frac{\frac{1}{n}\bar{X}_j^T\bar{X}_k}{\sqrt{\frac{1}{n}\bar{X}_j^T\bar{X}_j}\sqrt{\frac{1}{n}\bar{X}_k^T\bar{X}_k}} = \frac{\frac{1}{n}\bar{X}_j^T\bar{X}_k}{\frac{1}{n}\sqrt{\bar{X}_j^T\bar{X}_j}\sqrt{\bar{X}_k^T\bar{X}_k}} = \frac{\bar{X}_j^T\bar{X}_k}{\|\bar{X}_j\|\,\|\bar{X}_k\|} = \cos\theta$$

where  this is called the cosine similarity between the vectors in some contexts.

The sample covariance matrix is $\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix}$

by the formulas above, we see $\hat{\Sigma} = \frac{1}{n}\begin{pmatrix} \bar{X}_1^T\bar{X}_1 & \bar{X}_1^T\bar{X}_2 & \cdots & \bar{X}_1^T\bar{X}_d \\ \bar{X}_2^T\bar{X}_1 & \bar{X}_2^T\bar{X}_2 & \cdots & \bar{X}_2^T\bar{X}_d \\ \vdots & \vdots & & \vdots \\ \bar{X}_d^T\bar{X}_1 & \bar{X}_d^T\bar{X}_2 & \cdots & \bar{X}_d^T\bar{X}_d \end{pmatrix}$

$= \frac{1}{n}(\bar{D}^T\bar{D})$ ← sample scatter matrix

where $\bar{D} = D - \mathbf{1}\cdot\hat{\mu}^T = \begin{pmatrix} x_1^T - \hat{\mu}^T \\ \vdots \\ x_n^T - \hat{\mu}^T \end{pmatrix}$

The sample total variance is $tr(\hat{\Sigma}) = \hat{\sigma}_1^2 + \ldots + \hat{\sigma}_d^2$

The sample generalized variance is $|\hat{\Sigma}| = \det(\hat{\Sigma})$