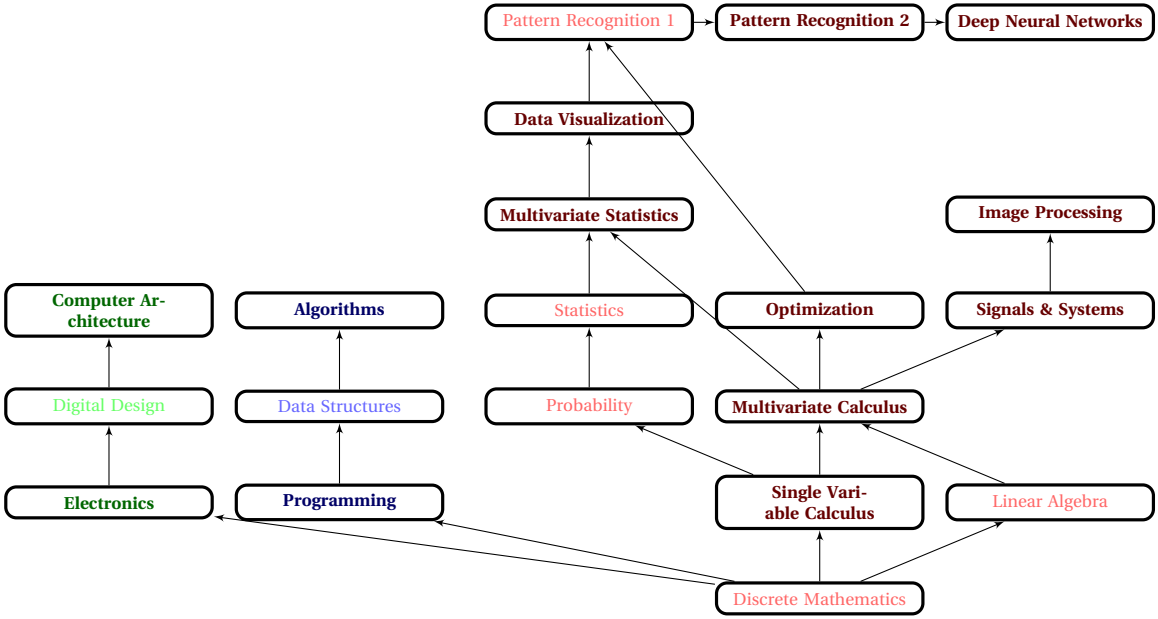
CS395
“Data Science”:
metro-rider snapshots from 15^{+} courses

Waleed A. Yousef, Ph.D.,

Human Computer Interaction Lab.,
Computer Science Department,
Faculty of Computers and Information,
Helwan University,
Egypt.

March 24, 2019

Anything new about the so-called "Data Science"?



Course Objectives

- To **direct, encourage, and push** students to study the whole set of 18-course-or-more field.
- To emphasize that is NOT a new field; it is a new fancy name.
- To provide a "trailer" of the field: (you study there not here)

Text

As **NOT** usual, lectures will **NOT** follow any particular text. However, some examples, figures, and numbers are borrowed from, e.g., Hastie et al. (2009); Duda et al. (2001); Rice (2007).

Prerequisites

Elementary probability and calculus (at the level of high school)

Contents

Chapter 1

Introduction

1.1 Examples from Real Life

Example 1 (comparison of two methods) *: Natrella M. (1963) "Experimental Statistics"; studies on latent heat of fusion of ice. The following table gives the change in total heat from ice at -72^{o}C to water at 0^{o}C in calories per gram of mass. (Example from (Rice, 2007, P. 423))*

Index	*Method-A*	*Method-B*
1	79.98	80.02
2	80.04	79.94
3	80.02	79.98
4	80.04	79.97
5	80.03	79.97
6	80.03	80.03
7	80.04	79.95
8	79.97	79.97
9	80.05	*NaN*
10	80.03	*NaN*
11	80.02	*NaN*
12	80	*NaN*
13	80.02	*NaN*

- *Are they different?*
- *Is it a coincidence?*
- *What is the probability of that?*

Example 2 (Email Spam) *:*

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`*.*

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

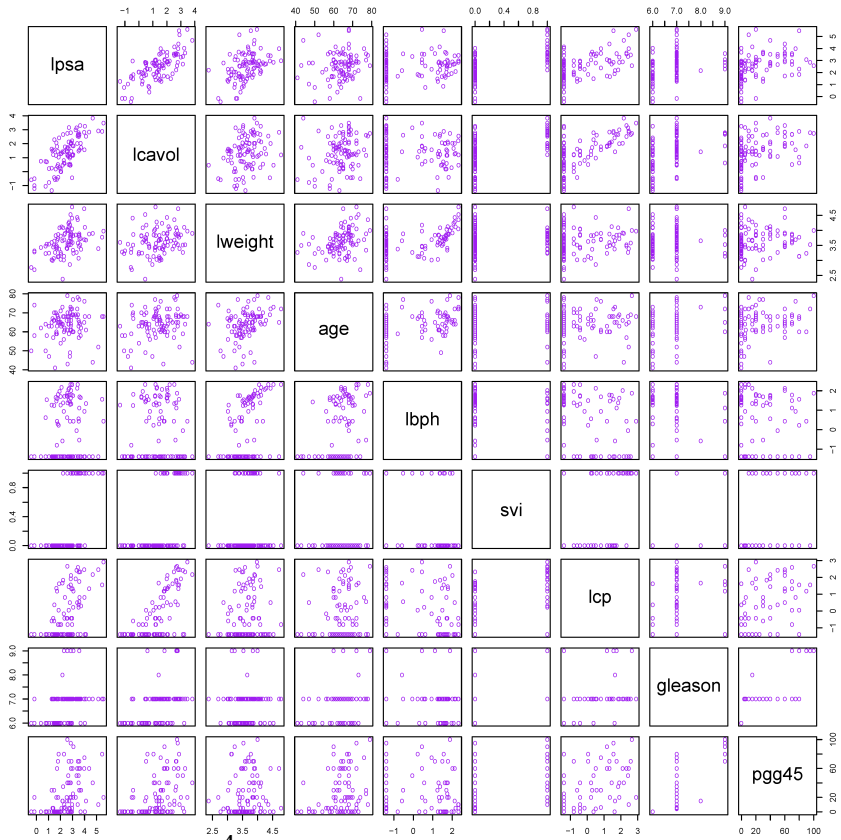
- *4601 email messages to try to predict whether the email was junk or not. The true outcome (`email` or `spam`) is available. This is also called classification problem (as will be explained later).*
- *The rule could be:*

  ```
  if (%george<0.6)&(%you>1.5) then spam
  else mail
  ```

- *But is this the "best" rule?*
- *Not all errors are equal!!*

Example 3 (Prostate Cancer) *:*

- *97 men (observations): predict the log of Prostate Specific Antigen (`lpsa`) from a number of measurements including log-cancer-volume (`lcavol`).*
- *This is a regression problem (of course supervised).*
- *Prostate-Specific Antigen (PSA): "Prostate-specific antigen, or PSA, is a protein produced by normal, as well as malignant, cells of the prostate gland."*



Example 4 (Handwritten Digit Recognition) *:*

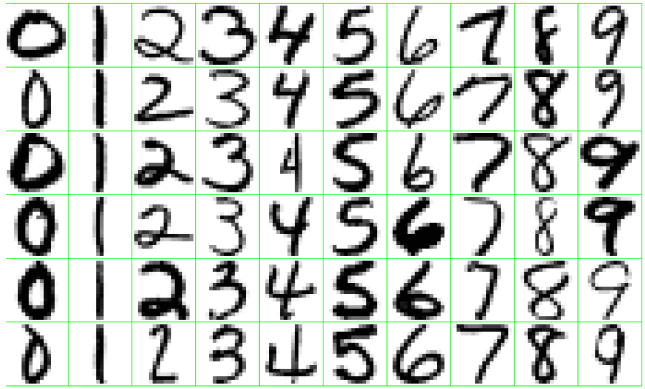


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

- *The data comes from the handwritten ZIP codes on envelops from U.S. postal mail.*
- *The images are* 16×16 *eight-bit grayscale maps, with each pixel ranging from 0-255.*
- *The task is to predict (classify) each image from its features (*16×16*) features to one of the digits.*
- *I'd like to see one of the projects to study this dataset and apply NN to it.*

Example 5 (DNA Expression Microarrays) *:*

- *DNA is the basic material that makes up human chromosomes. On the DNA chip, and through fluoroscopy, the gene-expression is measured.*
- *It ranges, e.g., between -6 to 6; positive values indicate higher expression (red) and negative indicate lower expression (green).*
- *Thousands of genes (features) exist; this is always ill-posed problem.*
- *The figure: experiment of 6830 genes (rows) (only 100 of them are displayed for clarity) and 64 samples (columns). The samples are 64 cancer tumor from different patients.*
- *Which samples are most similar to each other (across genes)?*
- *Which genes are most similar to each other (across samples)?*
- *Do certain genes show very high (or low) expression for certain cancer samples?*

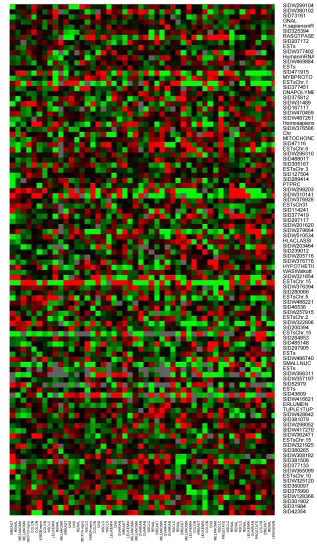


FIGURE 1.3. *DNA microarray data: expression ma-*

1.2 Ponder on Examples

- Elementary statistical analysis and understanding measurements and data is fundamental.
- "**Learning** is the process of estimating an unknown input-output dependency or structure of a system using a limited number of observations." (Cherkassky and Mulier, 1998).
- Statistical analysis and statistical learning plays a key role in many areas of science, finance and industry. Here are some examples of learning problems. **This is the newly called "Data Science"**

Chapter 2

A Snapshot of Probability

2.1 PDF, CDF, Inverse of CDF

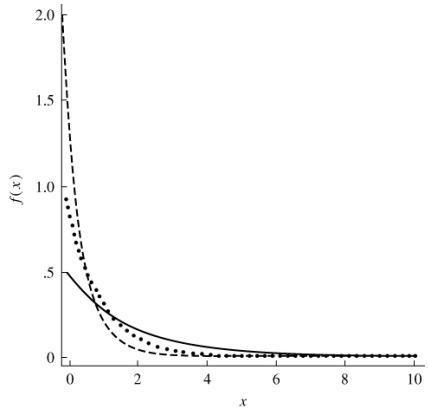
PDF:

$$P(a < X < b) = \int_a^b f(x)\ dx$$

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

$$P(X = c) = \int_c^c f(x)\ dx = 0,$$

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b)$$

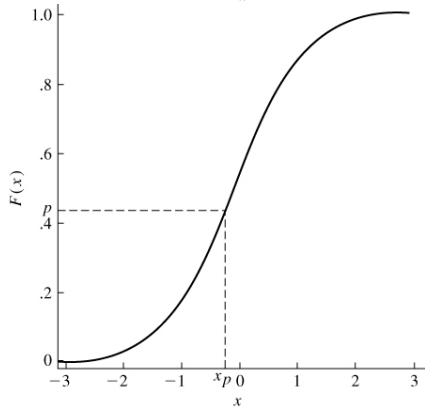


CDF:

$$F(x) = \int_{-\infty}^{x} f(u)\ du = P(X \le x)$$

$$f(x) = F'(x)$$

$$P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)\ dx.$$



Definition 6 (Inverse of CDF, F^{-1}) *: The p^{th} quantile is defined as, the value x_p of the r.v. that satisfies $F(x_p) = p$.*

- If F is monotonically (strictly) increasing, the pth quantile is unique (see figure).
- $F^{-1}(.5)$ is the median.
- $F^{-1}(.25)$ and $F^{-1}(.75)$is the lower and upper quartile.

Example 7 *Suppose*

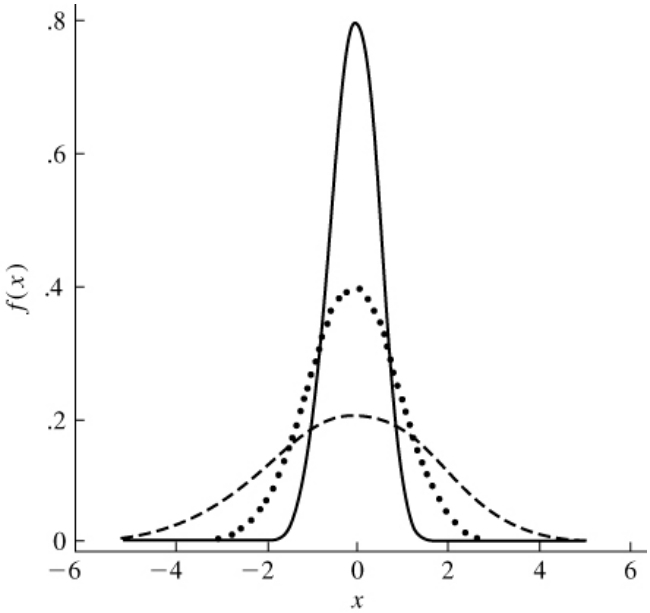
$$F(x) = x^2,\ 0 \le x \le 1,$$

$$\begin{aligned} x_p^2 &= p, \\ x_p &= \sqrt{p}, \\ x_{.5} &= \sqrt{.5} = .707 \\ x_{.25} &= \sqrt{.25} = .5 \\ x_{.75} &= \sqrt{.75} = .866 \end{aligned}$$

2.1.1 Normal Distribution: $Normal\left(\mu, \sigma^2\right)$ (our ever friend)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}, \ \sigma > 0.$$

$$\int_{-\infty}^{\infty} f(t)\ dt \stackrel{?}{=} 1$$

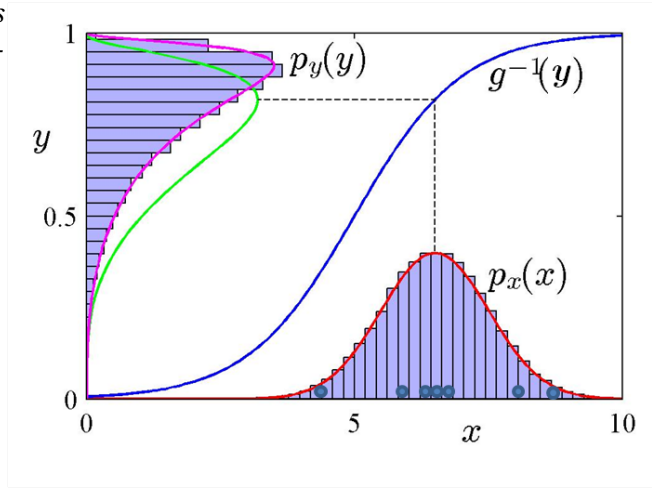


- $Normal$ because it is normal (statisticians)
- $Gaussian$ after Carl Friedrich Gauss in measuring errors (applied scientists)
- $Bell$ because it has a bell shape (some other parties)
- symmetric around μ
- no closed form CDF; called Φ.
- Again: repeating an experiment many times under this pdf, how data looks like? How this apply to next example?

2.1.2 Transformations

Theorem 8 (Transformation:) *If $Y = g(X)$, g is monotonically increasing (or decreasing) and g^{-1} exists then*

$$f_Y(y) = \left| \frac{1}{dy/dx} \right| f_X\left(g^{-1}(y)\right)$$



Special case:

$$Y = aX + b, \qquad X \sim N(\mu, \sigma^2)$$
$$Y \sim N\left((b + a\mu), (a\sigma)^2\right).$$

Interestingly: if

$$Z = \frac{X - \mu}{\sigma}$$
$$= \frac{1}{\sigma} X - \frac{\mu}{\sigma},$$

Then, Z has the standard Normal density:

$$Z \sim N(0, 1)$$

Corollary 9 *If $X \sim \mathcal{N}(\mu, \sigma)$ and $Z \sim \mathcal{N}(0,1)$ (a standard normal), then*

$$P(Z < z) = \int_{-\infty}^{z} f_Z(u)\, du = \Phi(z)$$

$$\Phi(z) = 1 - \Phi(-z)$$

$$\frac{(X-\mu)}{\sigma} \sim Z$$

$$P(X < x) = P\left(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = P\left(Z < \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Example 10 *[σ and μ]:*

$$\begin{aligned}
P\left(|X-\mu| < \sigma\right) &= P\left(-\sigma < X-\mu < \sigma\right) \\
&= P\left(-1 < \frac{X-\mu}{\sigma} < 1\right) \\
&= P\left(-1 < Z < 1\right) \\
&= \Phi\left(1\right) - \Phi\left(-1\right) \\
&= .68 \\
P\left(|X-\mu| < 2\sigma\right) &= \Phi\left(2\right) - \Phi\left(-2\right) \\
&= .9545, \\
P\left(|X-\mu| < 3\sigma\right) &= \Phi\left(3\right) - \Phi\left(-3\right) \\
&= .9973 \qquad \text{(almost all the probability measure)}
\end{aligned}$$

Example 11 (IQ test Scores X) :

- *Found that* $X \sim N\left(100, 15^2\right)$.
- *What is the probability* Pr *that* $X \in \left[120, 130\right]$*?*

$$\begin{aligned}
\Pr &= P\left(120 < X < 130\right) \\
&= P\left(\frac{120-100}{15} < \frac{X-100}{15} < \frac{130-100}{15}\right) \\
&= P\left(1.33 < Z < 2\right) \\
&= \Phi\left(2\right) - \Phi\left(1.33\right) \\
&= .9772 - .9082 \\
&= .069.
\end{aligned}$$

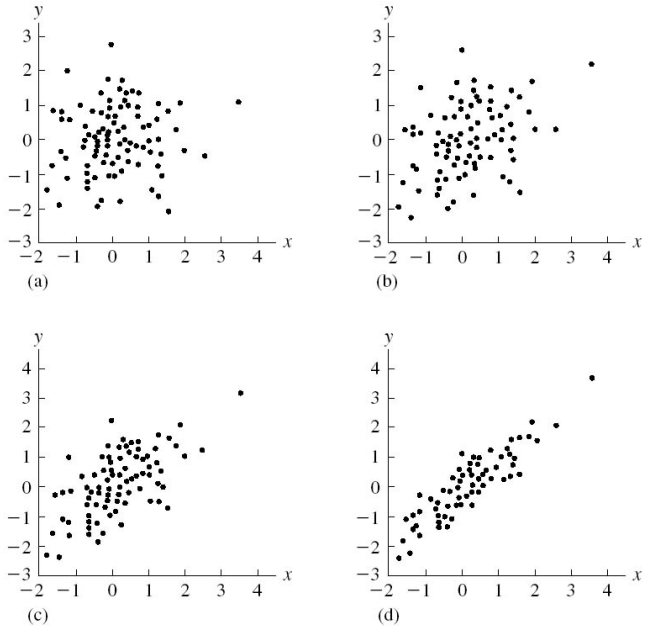
So, only 7% *of students takes grades in that range.*

2.2 Joint Distributions

- Number of predators and Number of preys for a particular species in ecology.
- Hight and Weight of particular category of distribution of people.
- A model for joint distribution of Age and Length in a population of fish.

Motivation by very simple example: $\{(0,0),(1,1)\}$ has different joint distribution than $\{(1,0),(0,1)\}$. However each of X and Y have the same marginal distribution.

Another example: The Height and Weight of some species of fish are reported for 100 fishes. How they are related together?



Third example: `lpsa` vs. `lcavol` (Ex. 3)

Definition 12 *If the joint cdf of two r.v. is differentiable, then the their joint pdf function is defined as*

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y);$$

and therefore

$$F_{XY}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(u,v)\, du\, dv.$$

and it can be also shown that (see extra material):

$$P((X,Y) \in A) = \int\int_{A} f_{XY}(u,v)\, du\, dv.$$

Marginal:

$$\begin{aligned} F_X(x) &= F_{XY}(x,\infty) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(u,v)\, du\, dv. \end{aligned}$$

Example 13 : *Consider the density*

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & 0 \le x \le y,\ \lambda > 0 \\ 0, & otherwise \end{cases}.$$

Take care, it can be re-written as

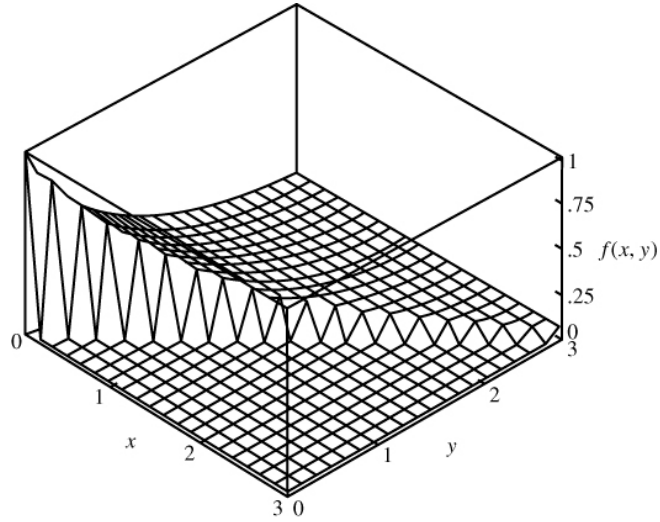
$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} I_{0\le x} I_{x\le y}, & \lambda > 0 \\ 0, & otherwise \end{cases}.$$

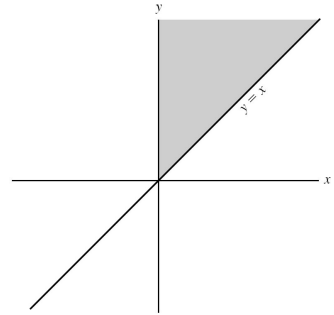
$$\begin{aligned} f_X(x) &= \int_x^\infty \lambda^2 e^{-\lambda y}\, dy \\ &= \lambda e^{-\lambda x},\ x \ge 0, \end{aligned}$$

which is $Exponential(\lambda)$

$$\begin{aligned} f_Y(y) &= \int_0^y \lambda^2 e^{-\lambda y}\, dx \\ &= \lambda^2 y e^{-\lambda y},\ 0 \le y, \end{aligned}$$

which is $Gamma(2,\lambda)$.





2.3 Independent R.V.s

If

$$F_{X_1X_2}(x_1,x_2) = F_{X_1}(x_1)F_{X_2}(x_2).$$

then

$$\begin{aligned} f_{X_1X_2}(x_1,x_2) &= \frac{\partial^2 \left[F_{X_1}(x_1)\,F_{X_2}(x_2)\right]}{\partial X_1 \partial X_2} \\ &= f_{X_1}(x_1)\,f_{X_2}(x_2), \end{aligned}$$

Also, if

$$f_{X_1X_2}(x_1,x_2) = f_{X_1}(x_1)\,f_{X_2}(x_2),$$

then

$$\begin{aligned} F_{X_1X_2}(x_1,x_2) &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1X_2}(x_1,x_2)\; dx_1\, dx_2 \\ &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1}(x_1)\,f_{X_2}(x_2)\; dx_1\, dx_2 \\ &= \int_{-\infty}^{x_1} f_{X_1}(x_1)\; dx_1 \int_{-\infty}^{x_2} f_{X_2}(x_2)\; dx_2 \\ &= F_{X_1}(x_1)\,F_{X_2}(x_2). \end{aligned}$$

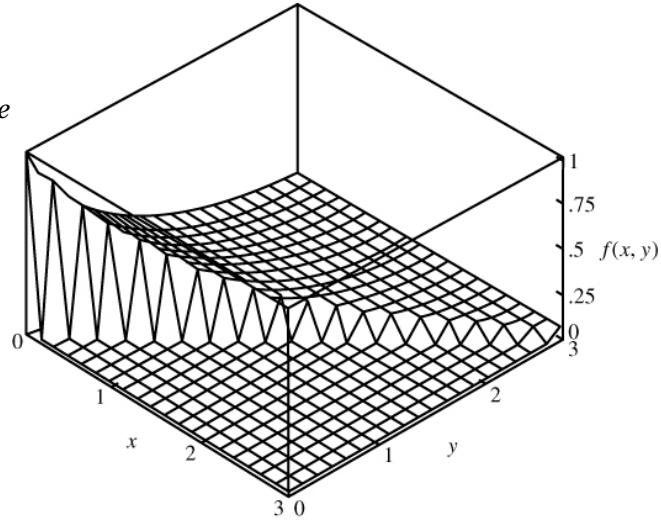
How this is reflected on data? let's discuss figure in Sec.2.2.

Example 14 (cont. Ex. 13) *:*

$$f_{XY}(x,y) = \begin{cases} \lambda^2 e^{-\lambda y}, & 0 \le x \le y,\ \lambda > 0 \\ 0, & otherwise \end{cases}$$

Take care, it looks like it factors; however it is not since

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} I_{0\le x} I_{x\le y}, & \lambda > 0 \\ 0, & otherwise \end{cases}.$$



The marginal was

$$f_X(x) = \lambda e^{-\lambda x},\ 0 \le x,$$
$$f_Y(y) = \lambda^2 y e^{-\lambda y},\ 0 \le y,$$

They are not independent, since

$$f_X f_Y \neq f_{XY}$$

2.4 Conditional Distributions

Definition 15 *The conditional density* $f_{Y|X}(y|x)$ *is defined as*

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)\, dy}$$

So the denominator is just a normalizing factor, and indeed $f_{Y|X}(y|x)$ is a pdf and integrates to one. **Let's try playing with Mathematica Notebook**

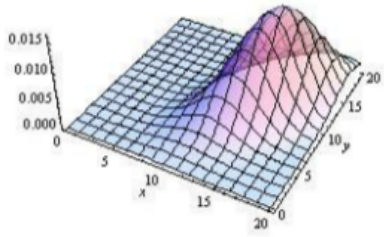
Law of total probability:

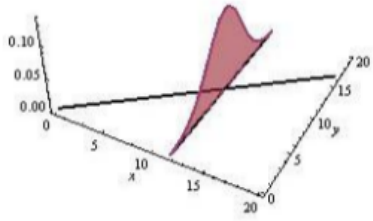
$$f_{XY}(x,y) = f_{Y|X}(y|x) f_X(x),$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\, dx,$$

Bayes' rule:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(x,y)\, dy} = \frac{f_{X|Y}(x|y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\, dy}$$





Example 16 (cont. Ex. 13) *:*

$$f(x,y)=\begin{cases}\lambda^2 e^{-\lambda y}, & 0\le x\le y,\ \lambda>0\\ 0, & otherwise\end{cases},$$

$$f_X(x)=\lambda e^{-\lambda x},\ 0\le x,$$

$$f_Y(y)=\lambda^2 y e^{-\lambda y},\ 0\le y.$$

$$f_{Y|X}(y|x)=\frac{\lambda^2 e^{-\lambda y}}{\lambda e^{-\lambda x}}$$

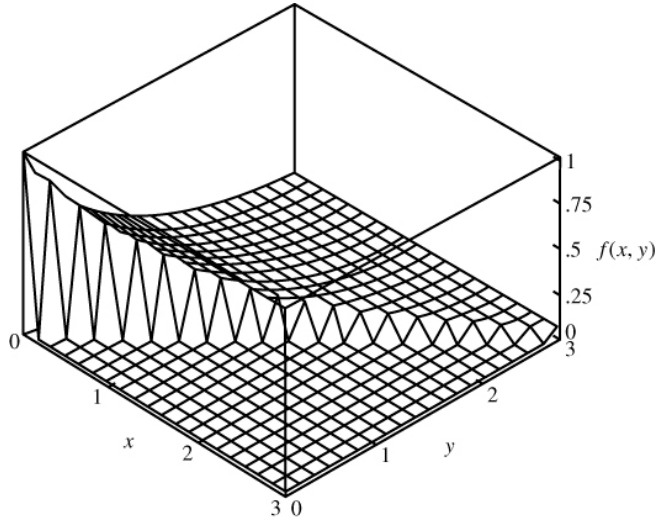
$$=\lambda e^{-\lambda(y-x)},\ 0\le x\le y,\ \lambda>0,$$

$$Y|X\sim Exponential(\lambda)$$

$$f_{X|Y}(x|y)=\frac{\lambda^2 e^{-\lambda y}}{\lambda^2 y e^{-\lambda y}}$$

$$=\frac{1}{y},\ 0\le x\le y,$$

$$X|Y\sim Uniform(0,1/y).$$



Notice that: we can generate (X,Y) *by generating* x*, followed by* $y|x$ *or by generating* y *followed* $x|y$

2.5 Expected Values

Definition 17 *If X is continuous r.v. then*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx,$$

If $\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty$; otherwise, it is undefined.

Example 18 ($Normal(\mu,\sigma^2)$) *:*

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

Substitute $z = x - \mu$

$$E(X) = \int_{-\infty}^{\infty} \underbrace{z \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z)^2}{2\sigma^2}\right]}_{symmetric} dz + \int_{-\infty}^{\infty} \underbrace{\mu \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z)^2}{2\sigma^2}\right]}_{pdf\ of\ Normal} dz = \mu.$$

So, the population parameters appear explicitly in the pdf. We say, $X \sim \mathcal{N}(\mu,\sigma^2)$. The figure shows the geometry of the normal distribution.

Data?

2.6 Variance and Standard Deviation

Definition 19 *If X is r.v with $E(X)$, then*

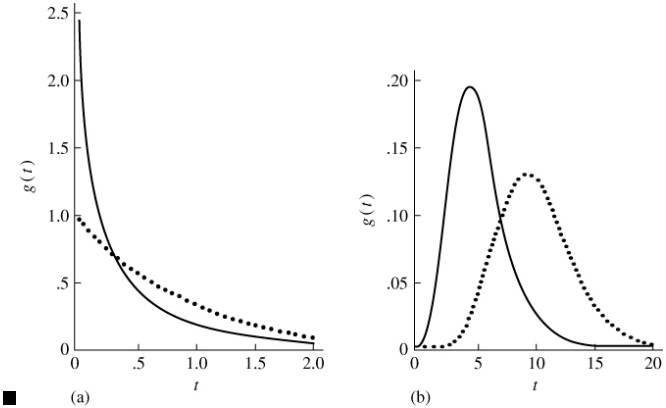
$$\sigma^2 \equiv \mathrm{Var}(X) = E\left[(X - E(X))^2\right],$$
$$\sigma \equiv \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)},$$

provided that $E\left[(X - E(X))^2\right] < \infty$.

Corollary 20 *The variance, if exists, can be given by*

$$\mathrm{Var}(X) = E\left(X^2\right) - [E(X)]^2$$

Proof is HW. ■



Intuition:

- we need some measure for "dispersion".
- SD has same units, so more meaningful.
- What is the pdf of $Y = X - E(X)$
- We could have defined it as:

$$\mathrm{Var}(X) = E(|X - E(X)|),$$

which is called absolute deviance.

How this is reflected on data?

Example 21 (Normal Distribution) *:*

$$\begin{aligned}\text{Var}(X) &= E\left(X-\mu\right)^2 \\ &= \int_{-\infty}^{\infty}\left(x-\mu\right)^2 \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right]dx = \sigma^2.\end{aligned}$$

Then, wonderful; the Normal pdf is expressed in terms of its population parameters

Discussion: revisit the 6-σ rule.

Theorem 22 *Suppose that* $\mathrm{Var}(X)$ *exists, and* X *is not necessarily normal; then*

1. *if* $Y = a + bX$ *then*
$$\mathrm{Var}(Y) = b^2 \mathrm{Var}(X) \qquad \mathrm{SD}(Y) = |b| \mathrm{SD}(X).$$
2. *if* $Y = \left(X - \mu_X\right)/\sigma_X$ *then*
$$E(Y) = 0, \qquad \mathrm{Var}(Y) = 1.$$

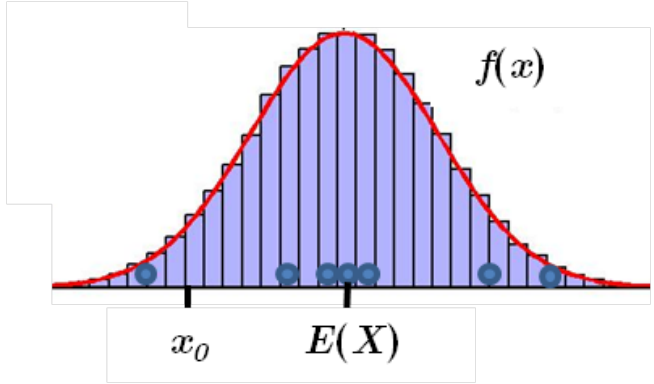
Proof is HW. ■

Example 23 (Simulation using Python.) *Simulate 3 different datasets, each is 100 observations, from:* $\mathcal{N}(0,16)$, $\mathcal{U}(0,4), \mathcal{E}xp(2)$. *Then plot the datasets with/without standardization:*

2.7 A Model for Measurement Error

- Suppose that we measure a constant x_0.
- Measurements are r.v. X, with μ and σ
- The error $X - x_0$ is a r.v; analyze it!
- Mean Squared Error (MSE):

$$MSE = E(X - x_0)^2$$



Theorem 24 (Mean Squared Error (MSE)) *:*

$$\begin{aligned} MSE &= Variance + Bias^2 \\ &= \sigma^2 + (\mu - x_0)^2 . \end{aligned}$$

Proof with common trick. :

$$\begin{aligned} MSE &= E(X - x_0)^2 \\ &= E\left((X - \mu) + (\mu - x_0)\right)^2 \\ &= E(X - \mu)^2 + 2(\mu - x_0)E(X - \mu) + (\mu - x_0)^2 \\ &= \sigma^2 + (\mu - x_0)^2 \end{aligned}$$

■

2.8 Covariance & Correlation

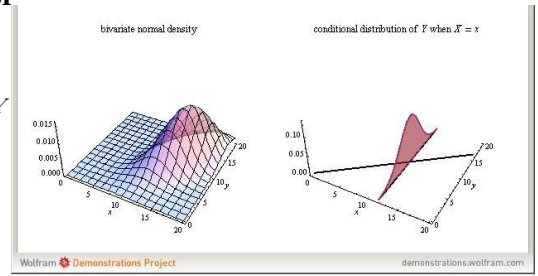
This section should have impact on your way of thinking and reading different situations in life

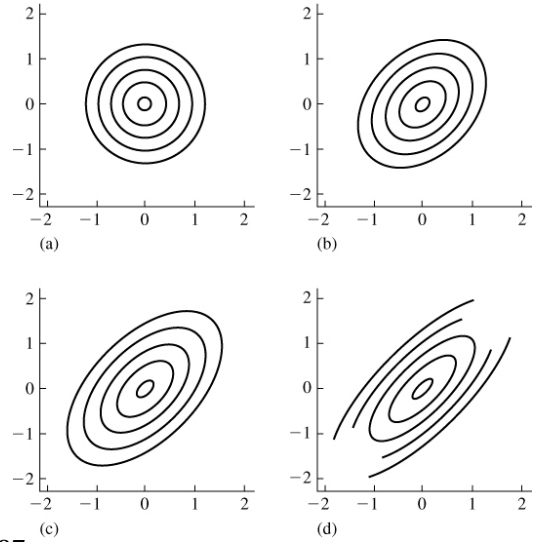
Definition 25 *If X and Y are two r.v. with μ_X and μ_Y*

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\left[(X-\mu_X)(Y-\mu_Y)\right]\\ &= E\left[(X-\mu_X)(Y-\mu_Y)\right]\\ &= E\left[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y\right]\\ &= E[XY]-\mu_X\mu_Y-\mu_X\mu_Y+\mu_X\mu_Y\\ &= E[XY]-\mu_X\mu_Y.\end{aligned}$$

Intuition:

- we need to measure "Association".
- $\mathrm{Cov}(X,Y)$ has the units of XY.
- If X and Y are independent $\mathrm{Cov}(X,Y)=0$





Example 26

$$f(x,y) = 2x + 2y - 4xy,\ 0 \le x, y \le 1,$$

$$f_X(x) = \int_0^1 (2x + 2y - 4xy)\, dy$$

$$= 1,\ 0 \le x \le 1$$

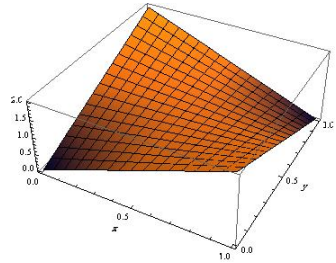
$$f_Y(y) = 1,\ 0 \le y \le 1$$

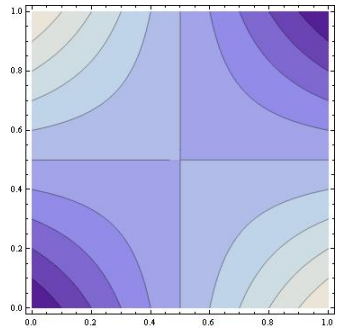
$$\mu_X = \mu_Y = \frac{1}{2}$$

$$\mathrm{Cov}(X,Y) = E(XY) - \mu_X \mu_Y$$

$$= \int_0^1 \int_0^1 xy(2x + 2y - 4xy)\, dx dy - \frac{1}{4}$$

$$= \frac{2}{9} - \frac{1}{4} = \frac{-1}{36}$$





Definition 27 (Correlation Coefficient) *: If X and Y are jointly distributed then:*

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Lemma 28 (Motivation for the definition) *:*

1. $-1 \le \rho_{XY} \le 1$ *and is dimensionless.*
2. *Under linear transformation* $U = a_0 + a_1 X$, *and* $V = b_0 + b_1 Y$*:*

$$\mathrm{Cov}(U,V) = a_1 b_1 \mathrm{Cov}(X,Y)$$

3. ρ_{XY} *is invariant under this linear transformation:*

Proof of 2 & 3.

$$\rho_{UV} = \frac{\mathrm{Cov}(a_0 + a_1 X, b_0 + b_1 Y)}{\sqrt{\mathrm{Var}(a_0 + a_1 X)\,\mathrm{Var}(b_0 + b_1 Y)}} = \frac{a_1 b_1 \mathrm{Cov}(X,Y)}{\sqrt{a_1^2 \mathrm{Var}(X)\, b_1^2 \mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \rho_{XY}$$

■

Example 29 (Revisit Ex. 26) *:*

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-1/36}{(1/\sqrt{12})(1/\sqrt{12})} = \frac{-1}{3}.$$

Theorem 30 $\rho = \pm 1$ *iff:* $P(Y = a + bX) = 1$ *for some* a, b.

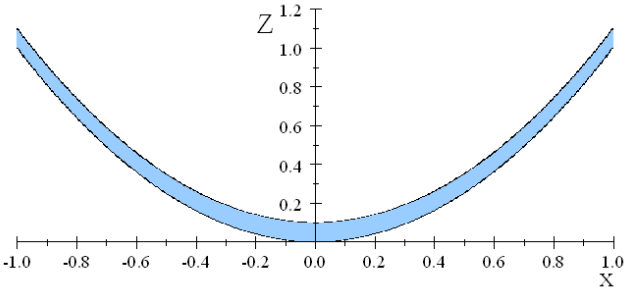
Example 31 (Counter Example) *:* X, Y *indep.,*

- $X \sim Uniform(-1,1)$; $f_X(x) = 1/2$.
- $Y \sim Uniform(0, 1/10)$; $f_Y(y) = 10$.
- $Z = X^2 + Y$: *what is* f_{XZ} *and* $\text{Cov}(X,Z)$*?*

$$f_{Z|X}(z|x) = 10,\ x^2 \le z \le x^2 + .1$$

$$f_{XZ}(x,z) = f_{Z|X}(z|x)\, f_X(x) = 10 \times \frac{1}{2} = 5.$$

$$= 5, -1 \le x \le 1, x^2 \le z \le x^2 + 0.1$$



$$\begin{aligned}\text{Cov}(X,Z) &= E[XZ] - E[X]E[Z] \\ &= E\left[X\left(X^2 + Y\right)\right] - 0E[Z] \\ &= E\left[X^3\right] + E[XY] \\ &= 0 + E[X]E[Y] \\ &= 0.\end{aligned}$$

Intuition: There is dependency, yet not linear.

Correlation is a measure of a linear relationship

X, Y	$\rho = 0$	$\rho \neq 0$
Dep.	*T*	*T*
Ind.	*T*	*F*

$$Independence \rightarrow \text{Cov} = 0$$

$$\text{Cov} \neq 0 \rightarrow Dependence.$$

Corollary 32 *Consider* $U = a_0 + \sum_{i=1}^{n} a_i X_i$

1. *In general:*

$$\operatorname{Var}(U) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{i>j} a_i a_j \operatorname{Cov}\left(X_i, X_j\right)$$

2. *If* X_i*s are uncorrelated (or independent):*

$$\operatorname{Var}(U) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i)$$

3. *If* X_i*s are i.i.d and* $a_i = 1$*:*

$$\operatorname{Var}(U) = n\sigma^2.$$

Observed Correlation Does Not Necessarily Imply Causation

May be one **or combination** of the following:

- **Example for A causes B:**

  Many

- **Example for B causes A:**

  Observation:

  (A): the more firemen fighting a fire

  (B): the bigger the fire is observed to be.

- **Example for (C) causes both:**

  Observation (Quinn et. al., 1999, Nature):

  (A): young children sleeping with the light

  (B): more likely to develop myopia

  Later study found that:

  infants sleeping with the light on caused the development of myopia!!

  However, they found that:

  parental myopia (C) is correlated with child myopia (B)

  myopic parents (C) were more likely to leave a light on (A) in their children's bedroom.

2.9 Conditional Expectation and Prediction

2.9.1 Definitions and Examples

Definition 33 (conditional expectation) *:*

$$E(Y|X=x)=\int y f_{Y|X}(y|x)\,dy \qquad (2.1)$$

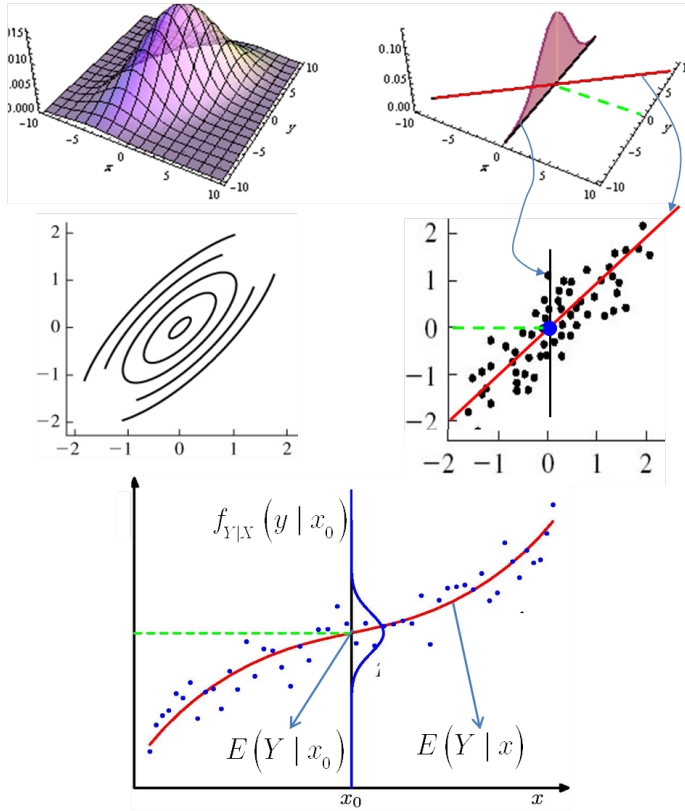
This takes us to "Regression Function":

Theorem 34 (Law of Total Expectation) *:* $E(Y)=E[E(Y|X)]$. *We can also write it as:*

$$E(Y)=E_X\left[E_{Y|X}(Y|X)\right].$$

Theorem 35 (Variance Decomposition) *:*

$$\mathrm{Var}[Y]=\underset{X}{\mathrm{Var}}\left[E_{Y|X}[Y|X]\right]+E_X\left[\underset{Y|X}{\mathrm{Var}}[Y|X]\right].$$



2.9.2 Prediction

Let's predict a r.v. by constant (Sec. 2.7)

$$\begin{aligned} MSE &= E(Y-c)^2 \\ &= \underbrace{\mathrm{Var}[Y]}_{\text{irreducible error}} + (E(Y)-c)^2 \\ c_{\min} &= \arg\min_c [MSE] \\ &= E(Y). \end{aligned}$$

If we replace E by $E_{Y|X}$

$$c = h(X) = E_{Y|X}(Y|X),$$

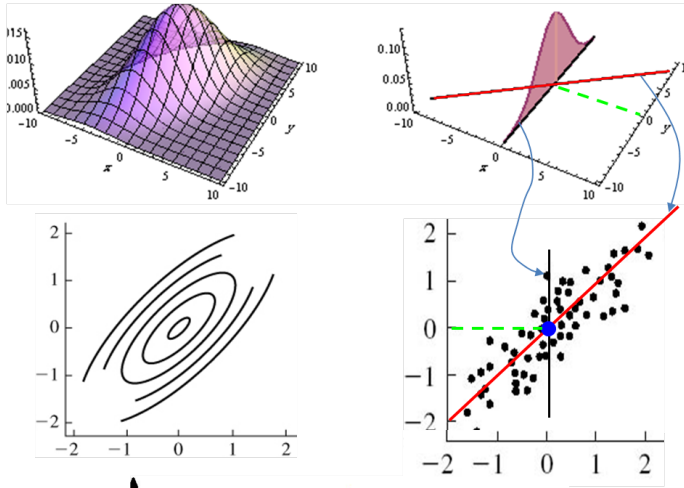
which minimizes

$$E_{Y|X}\left[(Y-h(X))^2|X\right],$$

and therefore minimizes

$$\begin{aligned} MSE &= E(Y-h(X))^2 \\ &= E_X E_{Y|X}\left[(Y-h(X))^2|X\right], \end{aligned}$$

which is the regression function. **This is what is Machine Learning is about!**



Chapter 3

A Snapshot of Statistics

3.1 Sampling, Statistics, and Weak Law of Large Numbers (WLLN)

Definition 36 (Sampling) *: The r.v. $X_1,\ldots,X_n$ are called a random sample of size n from the population F if $X_1,\ldots,X_n$ are i.i.d from F; and hence:*
$f_{X_1\ldots X_n}(x_1,\ldots,x_n)=\prod_i f(x_i)$.

Definition 37 (A statistic) *: Let $X_1,\ldots,X_n$ be a random sample of size n, and $T(x_1,\ldots,x_n)$ be a real- (or vector-) valued function whose domain includes the sample space of $(X_1,\ldots,X_n)$. Then the r.v. $Y=T(X_1,\ldots,X_n)$ is called a statistic.*

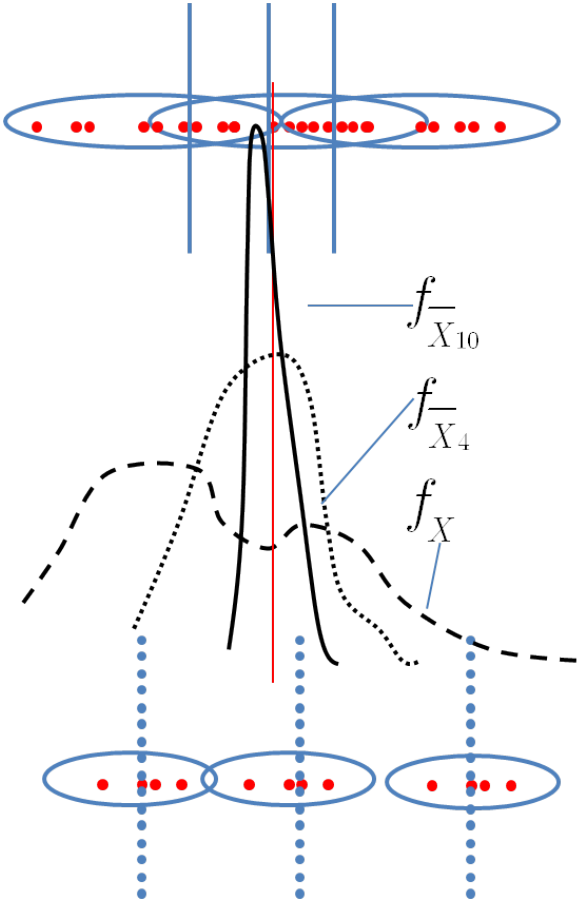
Definition 38 *The sample mean is a statistic defined as:*

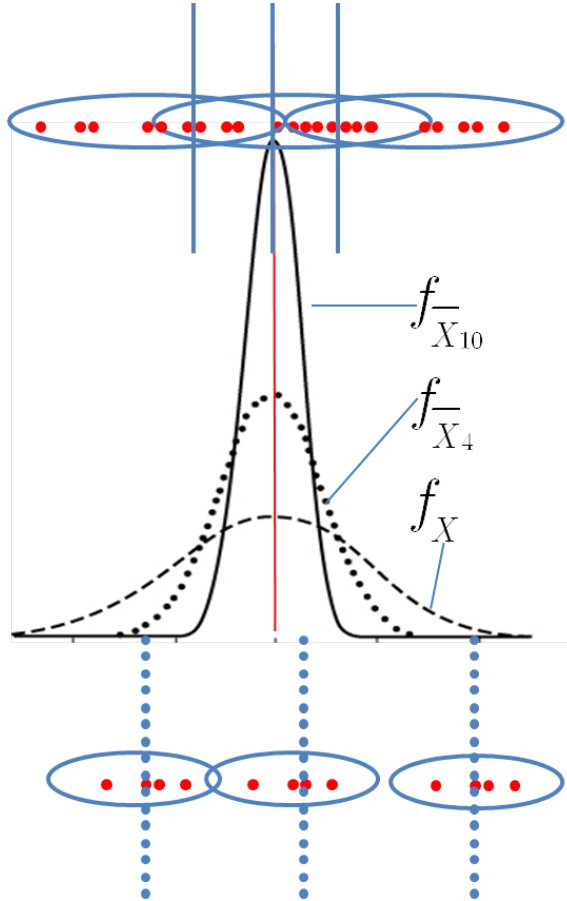
$$\overline{X}=\frac{1}{n}\sum_i X_i$$

		X_1	X_2	$\ldots$	X_n	$\overline{X}=\frac{1}{n}\sum_i X_i$
F	$Sample_1\longrightarrow$	$x_1,$	$x_2,$	$\ldots$	x_n	$\overline{x}=\frac{1}{n}\sum_i x_i$
F	$Sample_2\longrightarrow$	$x_1,$	$x_2,$	$\ldots$	x_n	$\overline{x}=\frac{1}{n}\sum_i x_i$
	$\vdots$					

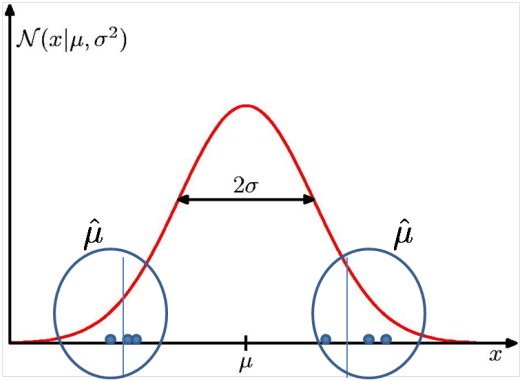
Theorem 39 (Weak Law of Large Numbers) *: If $X_1,\ldots,X_n$ is a s.r.v., independent with existing, and common, μ and σ^2 (but not necessarily identical) then $\overline{X}_n\left(=\frac{1}{n}\sum_{i=1}^n X_i\right)\xrightarrow{p}\mu$. A special case of the WLLN is when X_is are i.i.d. (repeated measurements)*

Let's see the meaning of the WLLN for $\overline{X}_n$:





3.2 Estimation and Estimators



Estimator is a real-valued function that tries to "close" in some sense to a population quantity. How "close"? Define a loss function, e.g., the Mean Square Error (MSE):

$$L\left(\widehat{\mu},\mu\right)=\left(\widehat{\mu}-\mu\right)^2.$$

And, define the Risk to be the Expected loss:

$$\mathrm{E}\left(\widehat{\mu}-\mu\right)^2$$

.

Important Decomposition for any estimator $\widehat{\mu}$:

$$\begin{aligned}\mathrm{E}\left(\widehat{\mu}-\mu\right)^2&=\mathrm{E}\left(\left(\widehat{\mu}-\mathrm{E}\widehat{\mu}\right)+\left(\mathrm{E}\widehat{\mu}-\mu\right)\right)^2\\&=\mathrm{E}\left(\widehat{\mu}-\mathrm{E}\widehat{\mu}\right)^2+\mathrm{E}\left(\mathrm{E}\widehat{\mu}-\mu\right)^2+2\mathrm{E}\left[\left(\widehat{\mu}-\mathrm{E}\widehat{\mu}\right)\left(\mathrm{E}\widehat{\mu}-\mu\right)\right]\\&=\mathrm{Var}\,\widehat{\mu}+Bias^2\left(\widehat{\mu}\right)\end{aligned}$$

We could have defined other loss functions, e.g., the absolute deviance loss:

$$L\left(\widehat{\mu},\mu\right)=\left|\widehat{\mu}-\mu\right|$$

One estimator may be better for one loss and not better for another loss.

3.3 Important Estimators

3.3.1 Estimation of μ_X

Sample mean $\overline{X}$ as an estimator of μ_X : $\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\mathrm{E}\,\overline{X} = \mathrm{E}\,\frac{1}{n}\sum_{i=1}^{n} x_i = \mathrm{E}\,X \;\; (=\mu)$$

$$Bias\left(\widehat{\mu}\right) = \mathrm{E}\,\widehat{\mu} - \mu = 0$$

$$\mathrm{Var}\,\widehat{\mu} = \frac{1}{n^2}\left[\sum_i \sigma^2 + \sum_i \sum_j \mathrm{Cov}\left(X_i, X_j\right)\right] = \frac{1}{n}\sigma^2$$

This means that from sample to sample it will vary with this variance.

An estimator with zero bias is called "unbiased". This means that on average it will be exactly as what we want.

HW: What about $\widehat{\mu} = X^{(1)}$.

3.3.2 Estimation of σ^2

$$
\begin{aligned}
\sigma^2 &= \mathrm{E}\left(X-\mu\right)^2 \\
&= \mathrm{E}\,X^2-\mu^2 \\
\widehat{\sigma^2} &= \frac{1}{n-1}\sum_i\left(x_i-\overline{X}\right)^2 \\
&= \frac{1}{n-1}\left(\sum_i x_i^2-n\overline{X}^2\right) \\
\mathrm{E}\widehat{\sigma^2} &= \sigma^2. \qquad \text{(unbiased)}
\end{aligned}
$$

3.3.3 Estimation of $\mathrm{Cov}(X,Y)$

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathrm{E}(X-\mu_X)(Y-\mu_Y) \\
&= \mathrm{E}\,XY - \mu_X\mu_Y \\
\widehat{\mathrm{Cov}}(X,Y) &= \frac{1}{n-1}\sum_i \left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right) \\
&= \frac{1}{n-1}\left(\sum_i x_i y_i - n\overline{XY}\right) \\
\mathrm{E}\,\widehat{\mathrm{Cov}}(X,Y) &= \mathrm{Cov}(X,Y) \qquad \text{(unbiased)}
\end{aligned}
$$

3.3.4 Quantile Estimation, Outliers, Cutoff, and Thick Tails

Recall the meaning of $F^{-1}(p)$; here, we will estimate them using order statistics. The ordered statistic $x_{(p=i.d)}$ is defined by interpolation as:

$$x_{(i.d)} = x_{(i)} + d(x_{(i+1)} - x_{(i)}) = (1-d)x_{(i)} + dx_{(i+1)} \qquad = \widehat{F}^{-1}(p) \quad \text{(sample } p^{\text{th}} \text{ quantile)}$$

$$x_{((n+1)/2)} = \begin{cases} x_{((n+1)/2)}, & n \text{ is odd.} \\ x_{(n/2+1/2)} = (1/2)(x_{(n/2)} + x_{(n/2+1)}) & n \text{ is even.} \end{cases} \qquad = \widehat{F}^{-1}(0.5) \quad \text{(sample median: M)}$$

$$x_{((1+(n+1)/2)/2)} = x_{((n+3)/4)} \qquad = \widehat{F}^{-1}(0.25)$$

(sample lower quartile: Q_L)

$$x_{(((n+1)/2+n)/2)} = x_{((3n+1)/4)} \qquad = \widehat{F}^{-1}(0.75)$$

(sample upper quartile: Q_U)

$$W_L = \min x_i \geq Q_L - 1.5(Q_U - Q_L) \qquad \text{(sample lower cutoff)}$$

$$W_U = \max x_i \leq Q_U + 1.5(Q_U - Q_L) \qquad \text{(sample upper cutoff)}$$

Example 40

34	*35*	*36*	*37*	*45*	*52*	*56*	*58*	*66*	*68*	*74*	*90*	*100*	*140*
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Rank of M, QL, QU is 7.5, 4.25, 10.75

$M = 56 + 0.5(58 - 56) = 57$

$Q_L = 37 + 0.25(45 - 37) = 39$

$Q_U = 68 + 0.75(74 - 68) = 72.5$

$d_Q = (72.5 - 39) = 33.5$

$Q_L - 1.5\, d_Q = 39 - 1.5 \times 33.5 = -11.25 \qquad W_L = 34$

$Q_U + 1.5\, d_Q = 72.5 + 1.5 \times 33.5 = 122.75 \qquad W_U = 100$

Example 41 (meaning of quantile from $X \sim \mathcal{N}(\mu,\sigma)$) *:*

$$
\begin{aligned}
p = F(x_p) = P(X < x_p) &= \Phi\left(\frac{x_p - \mu}{\sigma}\right) \\
F^{-1}(p) = x_p &= \mu + \left(\Phi^{-1}(p)\right)\sigma \\
Q_L = F^{-1}(0.25) &= \mu - 0.6745\sigma \\
Q_U = F^{-1}(0.75) &= \mu + 0.6745\sigma \\
d_Q &= 1.349\sigma \\
W_L = Q_L - 1.5d_Q &= \mu - 2.698\sigma \\
W_U = Q_U + 1.5d_Q &= \mu + 2.698\sigma \\
P(X < W_L) + P(X > W_U) = 2P(X < W_L) &= 2\Phi\left(\frac{(\mu - 2.698\sigma) - \mu}{\sigma}\right) = 2\Phi(-2.698) = 0.00698
\end{aligned}
$$

So, a sample (patch) of 1000 obs. will have almost 7 obs. outside the cutoffs.

Definition 42 (Hoaglin et al. (2000)) *:*
Outlier *observation with different underlying behavior as compared with the bulk of the data which deserves more investigation. The cutoffs W_L and W_U will be arbitrarily used for outlier detection. Outliers could be:*

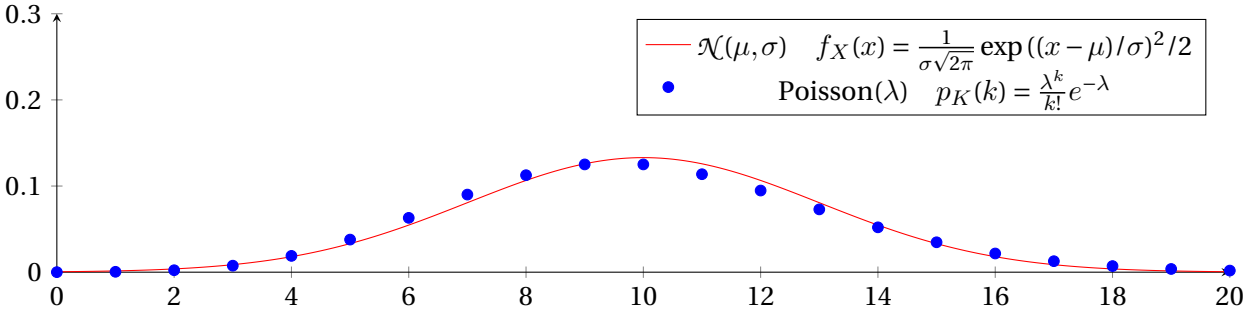
- *false value due to measurement error.*
- *right value due to thick tail.*

Resistance *insensitivity to misbehavior in data. A resistant method produces results that change only slightly when small part of the data is replaced by new numbers, possibly very different from the original ones.*

Chapter 4

A Snapshot of Data Visualization

4.1 A Quantitative Variable

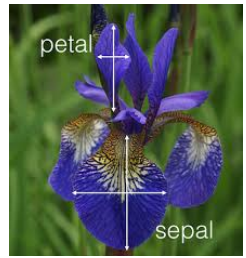


- Here, $\mu = 10,\ \sigma = 3, \lambda = 10$ (how do you know from figure?)
- $P(X = x) = 0, P(K = k) \neq 0$.
- How samples look like?
- What about cluttering (observations overlaying each other).

Example 43 (`iris` dataset) *: (150 observations, by R. A. Fisher, the father of Statistics)*

Index	*SepalLength*	*SepalWidth*	*PetalLength*	*PetalWidth*	*Class*
1	5.1	3.5	1.4	0.2	*Iris-setosa*
2	4.9	3	1.4	0.2	*Iris-setosa*
3	4.7	3.2	1.3	0.2	*Iris-setosa*
4	4.6	3.1	1.5	0.2	*Iris-setosa*
5	5	3.6	1.4	0.2	*Iris-setosa*
6	5.4	3.9	1.7	0.4	*Iris-setosa*
7	4.6	3.4	1.4	0.3	*Iris-setosa*
8	5	3.4	1.5	0.2	*Iris-setosa*
9	4.4	2.9	1.4	0.2	*Iris-setosa*
10	4.9	3.1	1.5	0.1	*Iris-setosa*
⋮					
51	7	3.2	4.7	1.4	*Iris-versicolor*
52	6.4	3.2	4.5	1.5	*Iris-versicolor*
53	6.9	3.1	4.9	1.5	*Iris-versicolor*
54	5.5	2.3	4	1.3	*Iris-versicolor*
55	6.5	2.8	4.6	1.5	*Iris-versicolor*
56	5.7	2.8	4.5	1.3	*Iris-versicolor*
57	6.3	3.3	4.7	1.6	*Iris-versicolor*
58	4.9	2.4	3.3	1	*Iris-versicolor*
59	6.6	2.9	4.6	1.3	*Iris-versicolor*
60	5.2	2.7	3.9	1.4	*Iris-versicolor*
⋮					
101	6.3	3.3	6	2.5	*Iris-virginica*
102	5.8	2.7	5.1	1.9	*Iris-virginica*
103	7.1	3	5.9	2.1	*Iris-virginica*
104	6.3	2.9	5.6	1.8	*Iris-virginica*
105	6.5	3	5.8	2.2	*Iris-virginica*
106	7.6	3	6.6	2.1	*Iris-virginica*
107	4.9	2.5	4.5	1.7	*Iris-virginica*
108	7.3	2.9	6.3	1.8	*Iris-virginica*
109	6.7	2.5	5.8	1.8	*Iris-virginica*
110	7.2	3.6	6.1	2.5	*Iris-virginica*
⋮					

- *Knowing the physics of the problem helps understanding data.*
- *Iris is a genus of species of flowering plants with showy flowers. (In Arabic: Alsawsan).*
- *Iris is extensively grown as ornamental plant, medicine, drugs.*



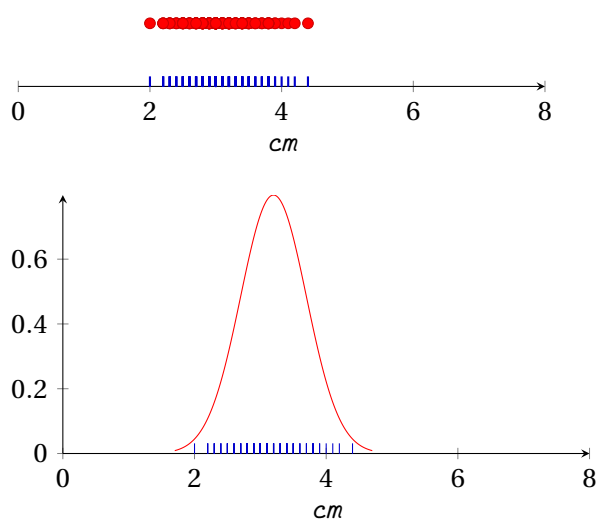
Per Hoaglin et al. (2000):

- How nearly symmetric the sample is?
- How spread out the numbers are?
- Whether a few values are far removed from the rest?
- Whether there are concentrations of data?
- Whether there are gaps in the data?

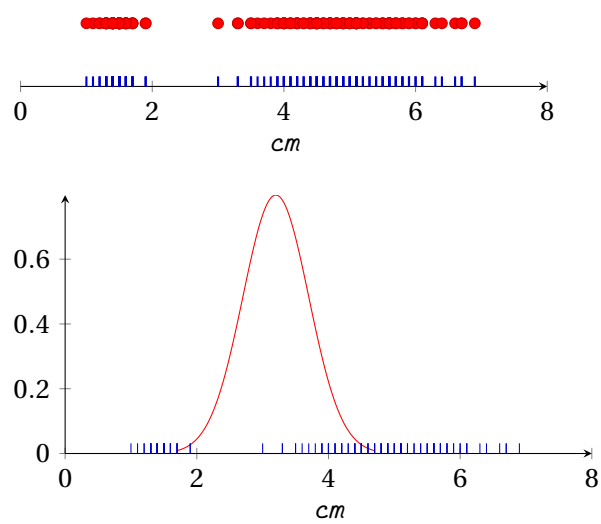
4.1.1 Rug Plot (the simplest ever)

Example 44 (`iris` **dataset**) .

SepalWidth



PetalLength



Hints for sense: Some observations clutter each other; standardize scale,

4.1.2 Histograms: (for more details check St 121.)

$$I_{(c)} = \begin{cases} 1 & \text{if } c \text{ is } T \\ 0 & \text{if } c \text{ is } F \end{cases}, \qquad \text{(indicator function)}$$

$$I_{(c)} \sim Bernoulli\left(\Pr(c)\right).$$

For data $x_1, \ldots, x_n$ divide the data range T to K equal regions of equal width Δ (so that $K = T/\Delta$)

$$\begin{aligned} T_k &= [t_0 + \Delta k, t_0 + \Delta(k+1)[\\ &= [t_k, t_{k+1}[,\ k = 0, \ldots, K-1, \end{aligned}$$

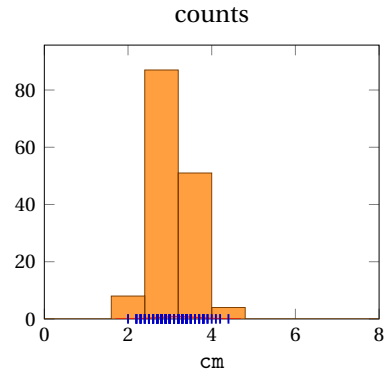
Notice: decreasing Δ increases K.
We have three versions of histogram:

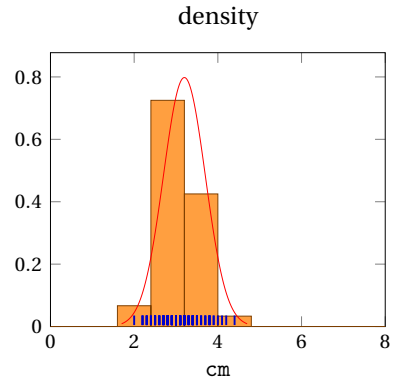
$$N_k = \sum_{i=1}^{n} I_{(X_i \in T_k)}, \qquad \text{(counts)}$$

$$R_k = \frac{N_k}{n} \xrightarrow{p} \Pr(X \in T_k) \qquad \text{(relative counts)}$$

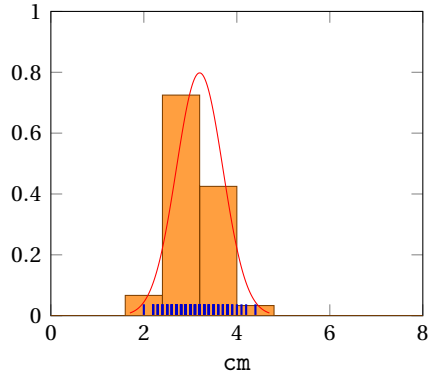
$$f_k = \frac{N_k}{\Delta n} \xrightarrow{p} \frac{\Pr(X \in T_k)}{\Delta} \approx \frac{f_X(t_k)\Delta}{\Delta} = f_X(t_k) \qquad \text{(density)}$$

SepalWidth

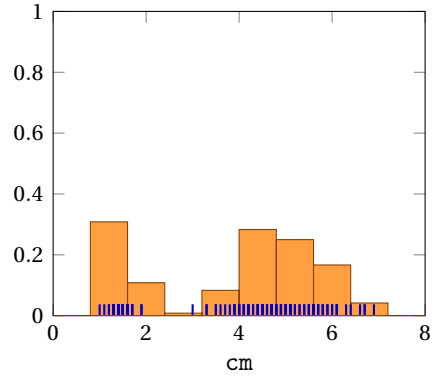


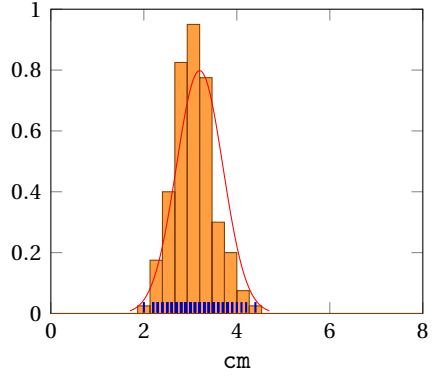


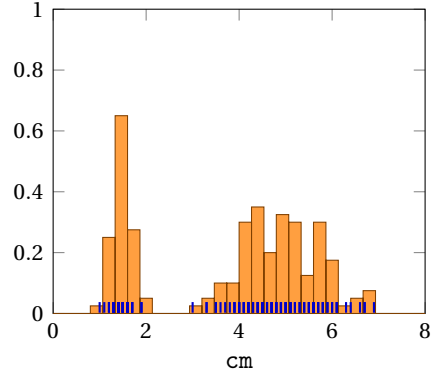
SepalWidth



PetalLength







Hints for sense: bins = 10 vs. 30; unify X and Y scale for comparison;

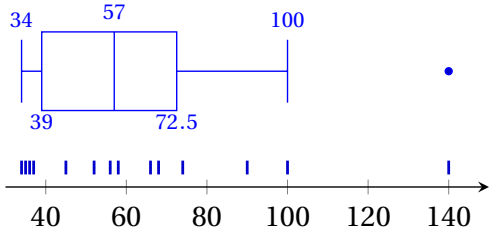
4.1.3 Box Plot

back to example 41

To observe a glance: location, spread, skewness, tail length, and outlying data points.

lower whisker	lower quartile	median	upper quartile	upper whisker
$Q_L - 1.5d_Q \le \min x_i = W_L$	Q_L	M	Q_U	$W_U = \max x_i \le Q_U + 1.5d_Q$

Example 45 (Letter Values) .



Rank of M, QL, QU is 7.5, 4.25, 10.75

$M = 56 + 0.5(58 - 56) = 57$

$Q_L = 37 + 0.25(45 - 37) = 39$

$Q_U = 68 + 0.75(74 - 68) = 72.5$

$d_Q = (72.5 - 39) = 33.5$

$39 - 1.5 \times 33.5 = -11.25$

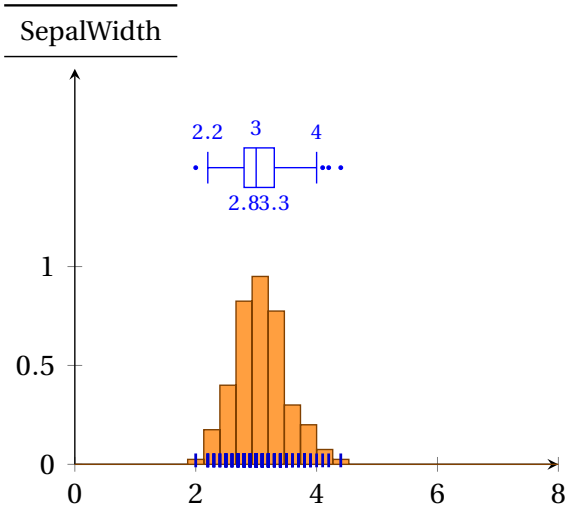
$72.5 + 1.5 \times 33.5 = 122.75$

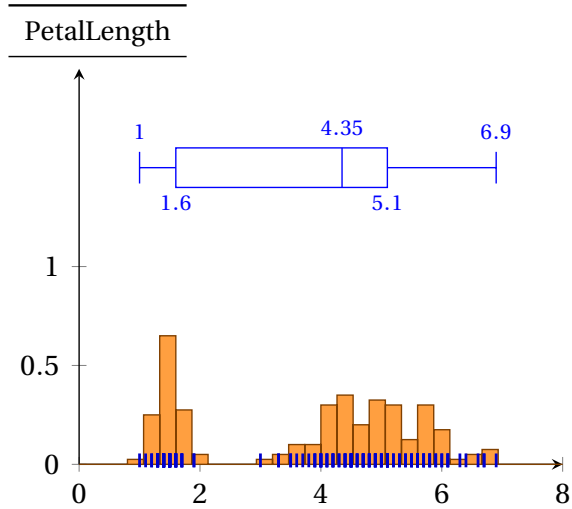
we could have defined a boxplot based on mean and variance => less resistant.

Why boxplot is not defined in terms of $W_L = \widehat{F}^{-1}(0.05)$

why boxplot is not defined in terms of mean and variance

for small patches $\underset{>}{\frac{\#\text{ of obs.}}{n}} \Pr(X < W_L)$





Comparison

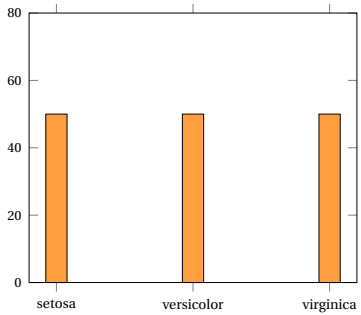
	Rug plot	**Histogram**	**Boxplot**
density	0 (clutter)	1	0 (region)
values	1	1	0 (region)
large N	0 (clutter)	1	1
resistance	0 (outliers)	0 (outliers)	1
discrete	0 (clutter)	1	1

Clutter could be alleviated by α-channel.

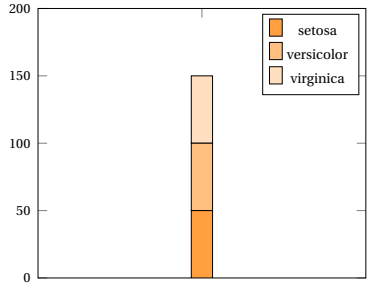
4.2 A Categorical Variable

Suppose we have only last column of table in Sec. 43; no numerical values. Only histogram-like charts: bar chart, stacked bar, pie chart, or any equivalent.

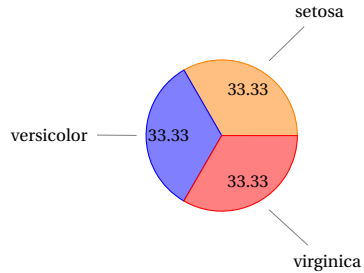
4.2.1 Bar chart



4.2.2 Stacked plot



4.2.3 Pie chart



- Bar chart is more professional and scientific; pie chart is more for illustration.
- More details can be put on the bar chart (including boxplot for each class, etc.)
- Bar chart and Stacked plot are utilized more for several patches.

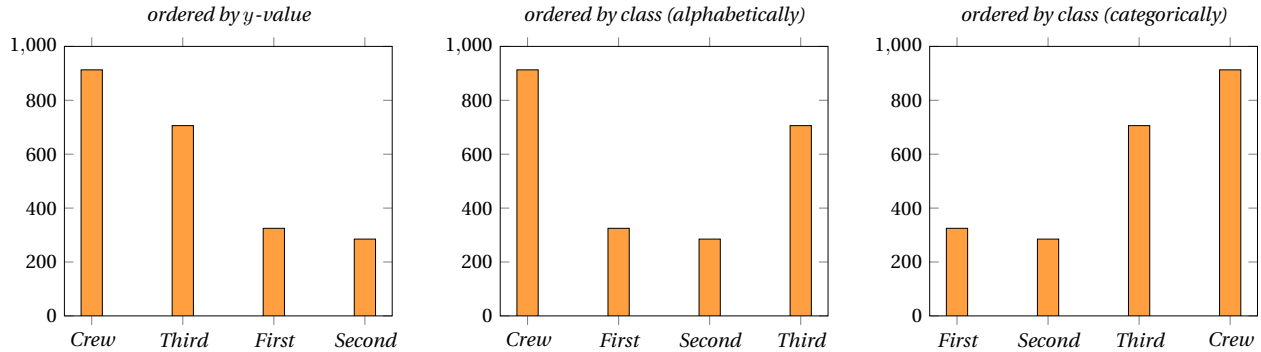
4.3 An Ordered Categorical Variable

Exactly as "bar chart" with ordered x-axis.

Example 46 (No. of Titanic passengers and crew) *: We can consider the variable (`passenger class`) as:*

- *categorical (as previous example) and order by y-value. (sorting will provide more information for the same ink).*
- *ordered categorical and order it alphabetically (nonsense in this example).*
- *ordered categorical and order it by class rank (makes sense here).*

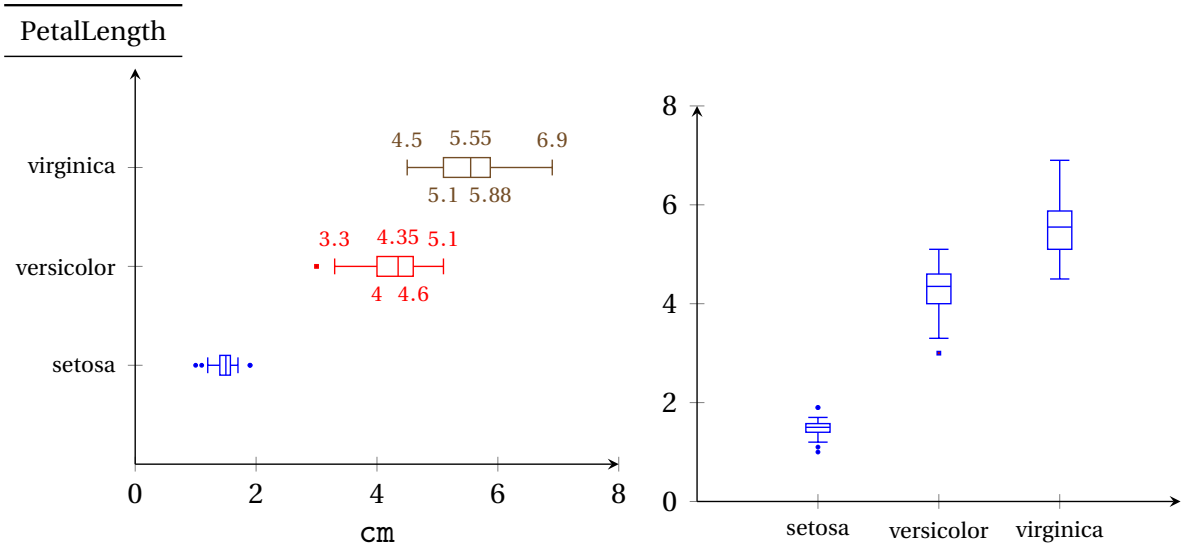
Reproduced



4.4 Two Variables

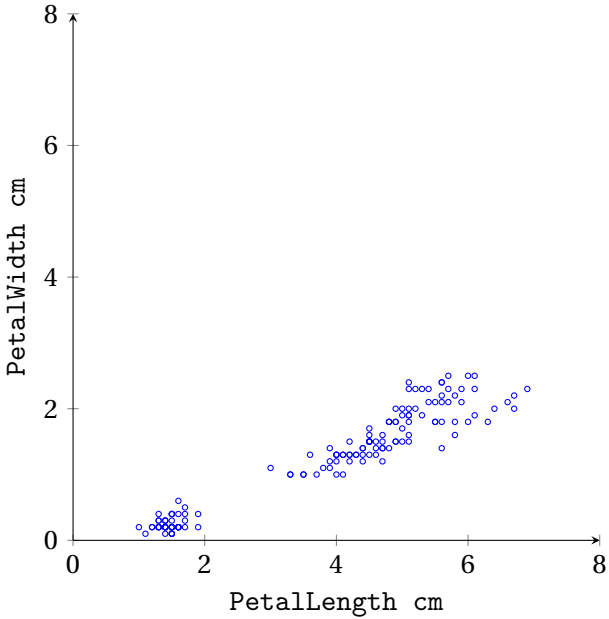
$$\{\texttt{Quant}, \texttt{Cat}\} \times \{\texttt{Quant}, \texttt{Cat}\}$$

4.4.1 Quantitative-Categorical

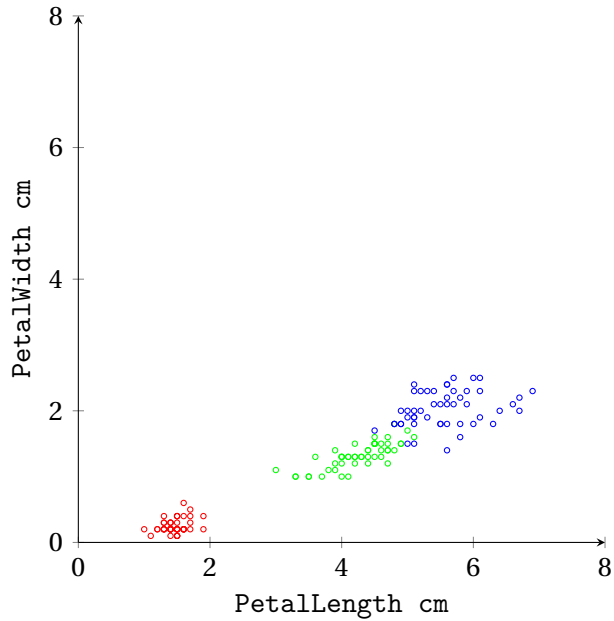


- coloring has no value here!
- categories here can be called patches.
- categories can be ordered by median

4.4.2 Quantitative-Quantitative



4.4.3 Quantitative-Quantitative-Categorical



- same scale for X and Y.
- DV points us towards the right model.

4.5 Contemplation in Higher Dimensions With a "Data Science" Coffee Blend: Mathematics, Software, الشعر عربى, and Pattern Recognition

4.5.1 Scatter Matrix Plot and Parallel Coordinates

4.5.2 Illustration vs. Exploration

- Objective: teaching vs learning
- Software: (LaTeXfamily) vs (data exploration family)

4.5.3 Rigor vs Ad-hoc & Theoreticians vs Practitioners

- Had we projected data to another subspace, would it be classified better!! Could we visualize it in a lower subspace, by using a set of orthonormal basis (Ex. 78)

$$\widehat{Y} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y \qquad \mathbf{P} = \frac{X_1 X_1'}{X_1' X_1} + \ldots + \frac{X_n X_n'}{X_n' X_n} = \mathbf{P}_1 + \ldots + \mathbf{P}_n, \qquad \widehat{Y} = \mathbf{P}_1 Y + \ldots \mathbf{P}_n Y.$$

- **Practitioners** should have a basic level of rigor when applying and **Researchers** should have the most rigorous possible level.

4.5.4 من قصيدة ``ذرف العبرات على من زعم التعرف على الأنماط بغير علوم الرياضيات"

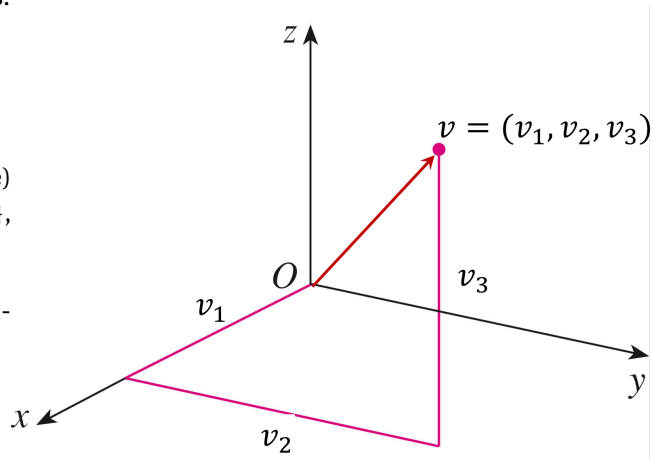
بياناتٍ رَسَمْناها فَكُنّ ..... بصائرَنا تُخَبِّرُنا اليَقينا
فَرُبَّ مُصَنِّفاتٍ قَد أَبَنّ ..... وَمَا غَفَلَتْ عُيونُ النَاظرينا

Chapter 5

A Snapshot of Linear Algebra

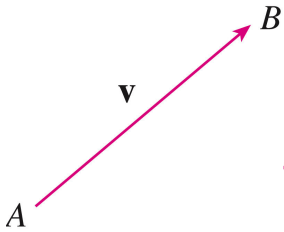
5.1 Back to School: visual space!

- We locate a point in a 3D space by three numbers.
- The coordinates are perpendicular.
- The order of the axes X , Y , Z: "right-hand" rule.
- The 3-tuple (3 ordered elements, or triple) $(v_1, v_2, v_3) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} = \mathbb{R}^3 = \{(x, y, z) | x, y, z \in \mathbb{R}\}$, the set of all points.
- The following are equivalent (some books differentiate; we do not):
  - the 3-tuple $v = (v_1, v_2, v_3)$.
  - the point $v = (v_1, v_2, v_3)$.
  - the arrow connecting O to v, i.e., the vector $v = \overrightarrow{Ov} = (v_1, v_2, v_3)$.
- The line segment $\overline{Ov}$ consists of **all** points, not only v.

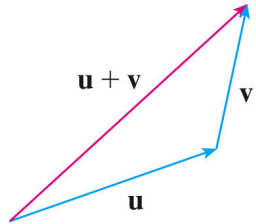


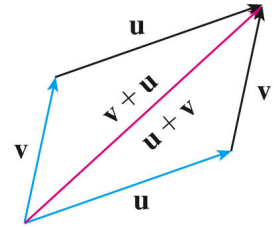
Definition 47 (Geometric Manipulation) .

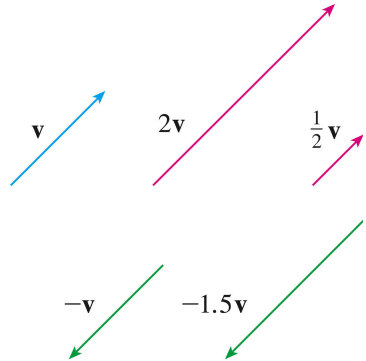
- *A vector is used to indicate a displacement in some direction; starting point is not important*
- *Start at any point A, move a distance in the direction of $\overrightarrow{Ov}$, and end at B. Then, $\overrightarrow{AB} = \overrightarrow{Ov} = v$. ($B \neq \overrightarrow{AB}$; but $v = \overrightarrow{Ov}$)*

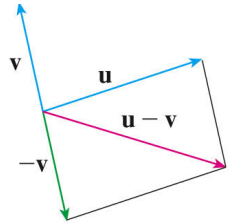


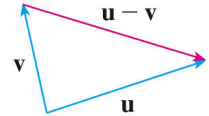
- *Addition: $u + v$*
- *Scalar Multiplication: If c is a scalar, then $u = cv$ is a vector whose length is $|c|\times$ length of v and direction:*
- *Scalar and Addition:*







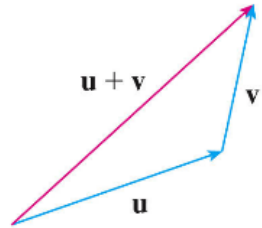




Definition 48 (Algebraic Treatment) . *Addition and Scalar: if* $a = (a_1, a_2, a_3)$, $b = (b_1, b_2, b_3)$*:*

$$\begin{aligned} a + b = &\quad (a_1 + b_1, a_2 + b_2, a_3 + b_3), \\ ca = &\quad (ca_1, ca_2, ca_3). \end{aligned}$$

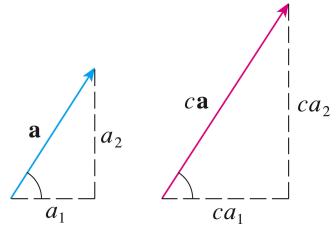
Proof of equivalence. Trivial. ■

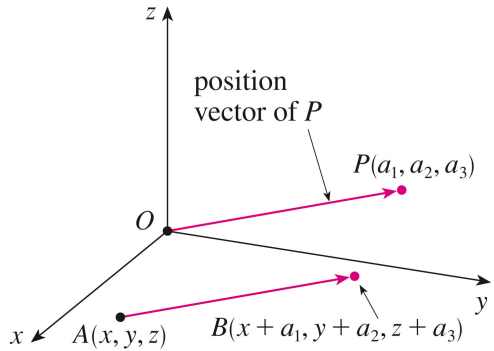


Lemma 49 (Properties of Vectors) . *For any two vectors* a *and* b*,*

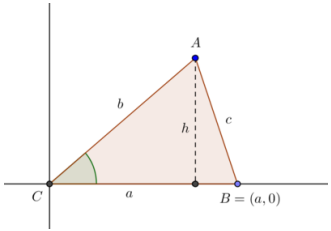
$$\begin{aligned} a + b &= b + a \\ a + (b + c) &= (a + b) + c \\ a + \mathbf{0} &= a \\ a + (-a) &= \mathbf{0} \\ c(a + b) &= ca + cb \\ (c + d)a &= ca + da \end{aligned}$$

Proof. It is quite straight forward to prove **(HW)** ■



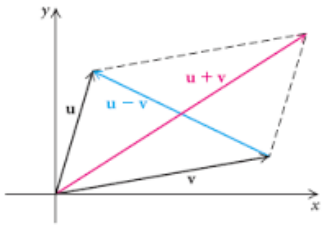


5.2 Angle, Lengths, and Dot Products (visual space and school again)



- Notation: the vector u, with 3-tuple (u_1, u_2, u_3) is written as: $u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ or $u' = (u_1, u_2, u_3)$.
- it is a school business to prove that (whether 2D or 3D):

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos\theta \\ \|u-v\|^2 &= \|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\cos\theta \\ 2\|u\|\|v\|\cos\theta &= (u_1^2+u_2^2) + (v_1^2+v_2^2) - (u_1-v_1)^2 - (u_2-v_2)^2 = 2u_1v_1 + 2u_2v_2 \\ \cos\theta &= \frac{u_1v_1+u_2v_2}{\|u\|\|v\|}. \end{aligned}$$



- This is why we defined the dot product to be: $u'v = u_1v_1 + u_2v_2 = \|u\|\|v\|\cos\theta$.
- When $u'v$ is zero we say they are orthogonal.
- If $u = v$, then $\theta = 0$, $u'u = u_1u_1 + u_2u_2 = \|u\|^2$.
- u is **unit vector** if $\|u\| = 1$. Then $\forall u$, $u/\|u\|$ is a unit vector.

$$\begin{aligned} u'v = \quad & \|v\|\|u\|\cos\theta = \|v\| \times \texttt{Projection Length of } u \texttt{ on } v \\ u'(v/\|v\|) = \quad & \texttt{Projection Length of } u \texttt{ on } v \\ v'(u/\|u\|) = \quad & \texttt{Projection Length of } v \texttt{ on } u. \end{aligned}$$

Lemma 50 (Properties) .

- *Basic properties:*

$$\begin{aligned} u'v &= v'u \\ \|au\| &= |a|\|u\| \\ a(u'v) &= (au)'v = au'v \\ (au+bv)'w &= au'w + bv'w \\ (u+v)'(u+v) &= u'u + 2u'v + v'v. \end{aligned}$$

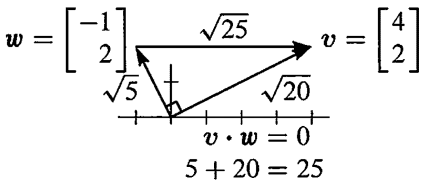
- *Cauchy-Shwartz inequality:* $-\|u\|\|v\| \le u'v \le \|u\|\|v\|$
  Proof. immediate from both: $-1 \le \cos\theta \le 1$ *and* $u'v = \|u\|\|v\|\cos\theta$. ■

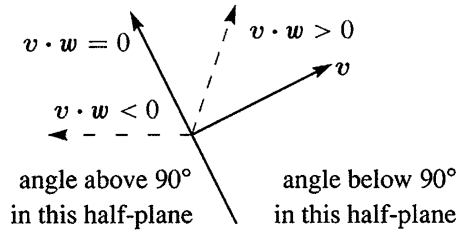
- *Traingular inequality:* $\|u+v\| \le \|u\| + \|v\|$
  Proof. $\|u+v\|^2 = (u+v)'(u+v) = u'u + 2u'v + v'v \le \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2$. ■

- *Then, we can generalize this definition in higher dimensions, and define the angle between two vectors for* $p > 3$.

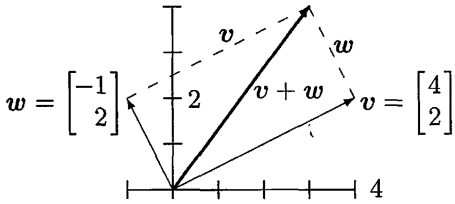
Example 51 $w = (-1,2)'$, $v = (4,2)'$, *then*

$$\cos\theta = \frac{w'v}{\|w\|\|v\|} = \frac{(-1)(4)+(2)(2)}{\sqrt{(-1)^2+(2)^2}\sqrt{(4)^2+(2)^2}} = \frac{0}{\sqrt{5}\sqrt{20}} = 0$$

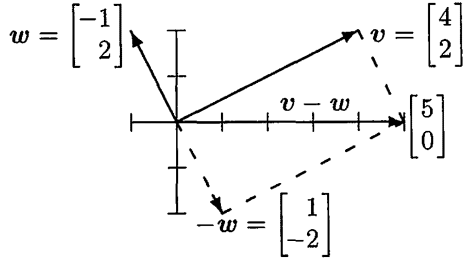




Example 52 (Linear Combination) .



$$v + w = \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$



$$v - w = \begin{bmatrix} 4 \\ 2 \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

5.3 Extension and Abstraction: Vectors and Linear Combinations

Extension in both: meaning and number of components to treat applications.

Definition 53 (Vector) *The ordered p-tuple $(v_1, v_2, \cdots, v_p)$, $v_i \in \mathcal{R}$, is called a p-dimensional vector.*

Definition 54 (dot product (inner product), length, angle)

$$\begin{aligned}\langle u, v\rangle = u.v = u'v &= \begin{pmatrix} u_1, & \cdots, & u_p \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \\ &= u_1v_1 + \cdots + u_pv_p = \sum_{i=1}^{p} u_iv_i \\ \|u\| = \sqrt{u'u} &= \sqrt{u_1^2 + u_2^2 + \cdots + u_p^2} \\ \cos\theta &= \frac{u'v}{\|u\|\|v\|}.\end{aligned}$$

Definition 55 (Linear Combination: generalization to adding vectors; this is the abstraction) .
Consider the two p-dimensional vectors v and w, and $c,\ d \in R$. We call $cv + dw$ a linear combination.

$$c\begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} + d\begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix} = \begin{pmatrix} cv_1 + dw_1 \\ \vdots \\ cv_p + dw_p \end{pmatrix}.$$

5.4 Rules for Matrix Operations

Definition 56 (Matrix) *: A matrix $A_{m\times n}$ is a square array (of size $m \times n$) of "objects" (could be numbers could be other blocks of matrices). The element a_{ij} is located in row i and column j respectively. We say $A = (a_{ij})$ or in some books $A = ((a_{ij}))$ to denote:*

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

- Languages handle matrices differently; e.g., Matlab images, C (row-wise), Fortran (column-wise), etc.
- Traversing matrices is $\Theta(m \times n)$.

5.4.1 Matrix Transpose

Definition 57 *The transpose of the matrix* $A_{m\times n}$ *is* $(A')_{n\times m}$ *, where* $A_{ij} = (A')_{ji}$

Example 58

$$A = \begin{pmatrix} 18 & 17 & 11 \\ 19 & -4 & 0 \end{pmatrix},\ A' = \begin{pmatrix} 18 & 19 \\ 17 & -4 \\ 11 & 0 \end{pmatrix}$$

Notice:

- $(A')' = A$.
- For vectors:

$$x = \begin{pmatrix} 19 \\ -4 \\ 0 \end{pmatrix},\ x' = \begin{pmatrix} 19 & -4 & 0 \end{pmatrix}.$$

  we usually write $x = \begin{pmatrix} 19 & -4 & 0 \end{pmatrix}'$, or $x' = \begin{pmatrix} 19 & -4 & 0 \end{pmatrix}$ to save vertical space.

Definition 59 (Symmetric Matrices (around diagonal)) *A square matrix* $A_{m\times m}$ *is called symmetric if* $A_{ij} = A_{ji}$ *; i.e.,* $A = A'$.

Example 60 (write a SW to check the symmetry of) $: A = \begin{pmatrix} 18 & 17 & 11 \\ 17 & -4 & 0 \\ 11 & 0 & 2 \end{pmatrix}$

5.4.2 Matrix Trace

Definition 61 *For a square matrix $A_{m\times m}$, the trace,* $\operatorname{trace}(A)$*, (for short* $\operatorname{tr}(A)$*), is defined as the sum of diagonal elements; i.e.,*

$$\operatorname{tr}(A) = \sum_{i=1}^{m} A_{ii}.$$

HW: write a C function to calculate the trace. (of course $\Theta(m)$*)*

Corollary 62

$$\operatorname{tr}(A) = \operatorname{tr}\left(A'\right).$$
$$\operatorname{tr}(x) = x \ \forall x \in \mathbb{R}.$$

Proof.

$$\operatorname{tr}(A) = \sum_i A_{ii} = \sum_i (A')_{ii} = \operatorname{tr}(\mathrm{A}').$$

■

Example 63

$$A = \begin{pmatrix} 1 & 7 & 6 \\ 8 & 3 & 9 \\ 4 & -2 & -8 \end{pmatrix} \Longrightarrow \operatorname{tr}(A) = -4.$$

5.4.3 Addition, Subtraction, and Scaling

Definition 64 *For equal size matrices $A_{m\times n}$ and $B_{m\times n}$, and for a scalar λ:*

- *the matrix $C = A \pm B$ is defined as*
$$C_{ij} = A_{ij} \pm B_{ij},$$
- *the matrix $D = \lambda A$ is defined as*
$$D_{ij} = \lambda A_{ij},$$
- *we say that $A = B$ if $A_{ij} = B_{ij}\ \forall i, j$.*
- *and a matrix, all of whose components are zeros, is written as $\mathbf{0}_{m\times n}$.*
- *Of course, $A + \mathbf{0} = A$*

Corollary 65 *It is quite easy to show that*

$$(A + B)' = A' + B'$$
$$\mathrm{tr}\,(A + B) = \mathrm{tr}\,(A) + \mathrm{tr}\,(B)$$

5.4.4 Matrix Multiplication

$$C = A_{m\times n}B_{n\times p} = \begin{pmatrix} a_{11} & & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & & a_{mn} \end{pmatrix}\begin{pmatrix} b_{11} & & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & & b_{np} \end{pmatrix} = C_{m\times p}$$

The general element C_{ij} is the dot product of Row_i and Col_j:

$$C_{ik} = a_i' b_k = \sum_{j=1}^{n} a_{ij}b_{jk} = a_{i1}b_{1k} + a_{i2}b_{2k} + \ldots + a_{in}b_{nk}.$$

However, we can partition either (or both) $A_{m\times n}$ and $B_{n\times p}$ as rows and/or columns to see the multiplication differently. This has a great value in mathematical treatments and semantics. We have only 4 ways to do that:

1. $A_{m\times 1}, B_{1\times p}$.
2. $A_{1\times n}, B_{n\times p}$.
3. $A_{1\times n}, B_{n\times 1}$.
4. $A_{m\times n}, B_{n\times 1}$.

Now, we will treat each case in detail.

1- As dot products

$$C = \begin{pmatrix} a_1' \\ \vdots \\ a_m' \end{pmatrix} \begin{pmatrix} b_1 & \cdots & b_p \end{pmatrix} \qquad (A_{m\times 1}B_{1\times p} \text{ partitioning})$$

$$= \begin{pmatrix} a_1'b_1 & & a_1'b_p \\ \vdots & \ddots & \vdots \\ a_m'b_1 & & a_m'b_p \end{pmatrix} \begin{pmatrix} \sum_{j=1}^n a_{1j}b_{j1} & & \sum_{j=1}^n a_{1j}b_{jp} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^n a_{mj}b_{j1} & & \sum_{j=1}^n a_{mj}b_{jp} \end{pmatrix}$$

$$C_{ik} = a_i'b_k = \sum_{j=1}^{n} a_{ij}b_{jk} = a_{i1}b_{1k} + \cdots + a_{in}b_{nk}$$

Example 66

$$= \begin{pmatrix} 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 3 & 2 & 0 \\ 1 & 4 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 3+4 & 2+16 & 0-4 \\ 3+5 & 2+20 & 0-5 \end{pmatrix}$$

$$= \begin{pmatrix} 7 & 18 & -4 \\ 8 & 22 & -5 \end{pmatrix}$$

2- As linear combinations of columns of A

$$C = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_{11} & & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & & b_{np} \end{pmatrix} \qquad (A_{1\times n}B_{n\times p} \text{ partitioning})$$

$$= \begin{pmatrix} b_{11}a_1 + \cdots + b_{n1}a_n & \cdots & b_{1p}a_1 + \cdots + b_{np}a_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_j b_{j1}a_j & \cdots & \sum_j b_{jp}a_j \end{pmatrix}$$

$$= \begin{pmatrix} c_1 & \cdots & c_p \end{pmatrix}$$

$$C_{ik} = (c_k)_i = \left(\sum_j b_{jk}a_j\right)_i = \sum_j (b_{jk}a_j)_i = \sum_j b_{jk}a_{ij}.$$

Example 67

$$C = \begin{pmatrix} 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 3 & 2 & 0 \\ 1 & 4 & -1 \end{pmatrix}$$

$$= \left(3\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1\begin{pmatrix} 4 \\ 5 \end{pmatrix} \quad 2\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 4\begin{pmatrix} 4 \\ 5 \end{pmatrix} \quad 0\begin{pmatrix} 1 \\ 1 \end{pmatrix} + -1\begin{pmatrix} 4 \\ 5 \end{pmatrix} \right)$$

$$= \left(\begin{pmatrix} 3+4 \\ 3+5 \end{pmatrix} \quad \begin{pmatrix} 2+16 \\ 2+20 \end{pmatrix} \quad \begin{pmatrix} 0-4 \\ 0-5 \end{pmatrix} \right) = \begin{pmatrix} 7 & 18 & -4 \\ 8 & 22 & -5 \end{pmatrix}$$

3- As linear combinations of rows of B

$$C=\begin{pmatrix} a_{11} & & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & & a_{mn} \end{pmatrix}\begin{pmatrix} b'_1 \\ \vdots \\ b'_n \end{pmatrix} \quad (A_{m\times n}B_{n\times 1} \text{ partitioning})$$

$$=\begin{pmatrix} a_{11}b'_1+\cdots+a_{1n}b'_n \\ \vdots \\ a_{m1}b'_1+\cdots+a_{nm}b'_n \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{j=1}^{n} a_{1j}b'_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}b'_j \end{pmatrix}$$

$$=\begin{pmatrix} c'_1 \\ \vdots \\ c'_m \end{pmatrix}$$

$$C_{ik}=(c'_i)_k=\Big(\sum_j a_{ij}b'_j\Big)_k=\sum_j\big(a_{ij}b'_j\big)_k=\sum_j a_{ij}b_{jk}.$$

Example 68

$$=\begin{pmatrix}1 & 4\\ 1 & 5\end{pmatrix}\begin{pmatrix}3 & 2 & 0\\ 1 & 4 & -1\end{pmatrix}$$

$$=\begin{pmatrix}1\begin{pmatrix}3 & 2 & 0\end{pmatrix}+4\begin{pmatrix}1 & 4 & -1\end{pmatrix}\\ 1\begin{pmatrix}3 & 2 & 0\end{pmatrix}+5\begin{pmatrix}1 & 4 & -1\end{pmatrix}\end{pmatrix}$$

$$=\begin{pmatrix}\begin{pmatrix}3+4 & 2+16 & 0-4\end{pmatrix}\\ \begin{pmatrix}3+5 & 2+20 & 0-5\end{pmatrix}\end{pmatrix}=\begin{pmatrix}7 & 18 & -4\\ 8 & 22 & -5\end{pmatrix}$$

4- As summation of outer products, each is a matrix

$$C = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1' \\ \vdots \\ b_n' \end{pmatrix} \quad (A_{1\times n}B_{n\times 1} \text{ partitioning})$$

$$= a_1 b_1' + \cdots + a_n b_n'$$

$$= \sum_{j=1}^{n} a_j b_j',$$

$$a_j b_j' = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix} b_j' = \begin{pmatrix} a_{1j} b_j' \\ \vdots \\ a_{nj} b_j' \end{pmatrix},$$

$$C_{ik} = \Big(\sum_j a_j b_j'\Big)_{ik} = \sum_j \big(a_j b_j'\big)_{ik} = \sum_j a_{ij} b_{jk}.$$

Example 69

$$= \begin{pmatrix} 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} 3 & 2 & 0 \\ 1 & 4 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & 0 \end{pmatrix} + \begin{pmatrix} 4 \\ 5 \end{pmatrix} \begin{pmatrix} 1 & 4 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 3 & 2 & 0 \\ 3 & 2 & 0 \end{pmatrix} + \begin{pmatrix} 4 & 16 & -4 \\ 5 & 20 & -5 \end{pmatrix} = \begin{pmatrix} 7 & 18 & -4 \\ 8 & 22 & -5 \end{pmatrix}$$

Product with Diagonal Matrix

Definition 70 *A matrix D is diagonal if $D_{ij} = 0 \; \forall i \neq j$; i.e.,*

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & d_m \end{pmatrix}.$$

Since there is no confusion, we subscript d_i instead of d_{ii}. We also, for short, write $D = \text{diag}(d_1, \ldots, d_n)$

Row scaling:

$$D_{m\times m} A_{m\times n} = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_m \end{pmatrix} \begin{pmatrix} a'_1 \\ \vdots \\ a'_m \end{pmatrix} = \begin{pmatrix} d_1 a'_1 \\ \vdots \\ d_m a'_m \end{pmatrix}$$

Column scaling:

$$A_{m\times n} D_{n\times n} = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} d_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_n \end{pmatrix} = \begin{pmatrix} a_1 d_1 & \cdots & a_n d_n \end{pmatrix}$$

Definition 71 *The identity matrix I is a special case diagonal matrix and defined as*

$$I_{m\times m} = \text{diag}\begin{pmatrix} 1, \ldots, 1 \end{pmatrix}$$

It is obvious that $IA = AI = A$.

Transpose of a Product

Lemma 72 *For conforming matrices $A_{m\times n}$ and $B_{n\times p}$,*

$$(AB)' = B'A',$$

and more general

$$(A_1 \cdots A_n)' = A_n' \cdots A_1'.$$

Trace of a Product

The trace is defined only for a square matrix; hence, for a product to have a trace it must be $A_{m\times n}B_{n\times m}$.

Lemma 73 *For two-side conforming matrices $A_{m\times n}$ and $B_{n\times m}$,*

$$\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)\,,$$

and more general

$$\mathrm{tr}\,(A_1 \cdots A_n) = \mathrm{tr}\,(A_n \cdots A_1)\,.$$

5.4.5 The Laws of Algebra

Theorem 74 $\forall A_{m\times n},\ B_{m\times n},\ C_{m\times n}, c$ *scalar, we have*

$$A+B=B+A \qquad \text{(commulative)}$$
$$c(A+B)=cA+cB \qquad \text{(distributive)}$$
$$A+(B+C)=(A+B)+C, \qquad \text{(associative)}$$

and

$$C(A+B)=CA+CB, \qquad (\forall A_{m\times n}, B_{m\times n}, C_{k\times m})$$
$$(A+B)C=AC+BC, \qquad (\forall A_{m\times n}, B_{m\times n}, C_{n\times p})$$
$$A(BC)=(AB)C \qquad (\forall A_{m\times n}, B_{n\times p}, C_{p\times q})$$
$$A_{m\times n}B_{n\times m}\neq B_{n\times m}A_{m\times n}$$
$$A_{m\times m}B_{m\times m}=\neq B_{m\times m}A_{m\times m}$$

Example 75 (Counter Example for $AB\neq BA$)

$$\begin{pmatrix}1&2\\3&4\end{pmatrix}\begin{pmatrix}0&2\\3&3\end{pmatrix}=\begin{pmatrix}6&8\\12&18\end{pmatrix}=\begin{pmatrix}0&2\\3&3\end{pmatrix}\begin{pmatrix}1&2\\3&4\end{pmatrix}$$

$$\underset{A}{\begin{pmatrix}1&2\\3&4\end{pmatrix}}\underset{B}{\begin{pmatrix}0&1\\3&5\end{pmatrix}}=\begin{pmatrix}6&11\\12&23\end{pmatrix}\neq\begin{pmatrix}3&4\\18&26\end{pmatrix}=\underset{B}{\begin{pmatrix}0&1\\3&5\end{pmatrix}}\underset{A}{\begin{pmatrix}1&2\\3&4\end{pmatrix}}$$

Example 76 *Factor* $Y = XPX + QX^2 + X$ *and find the constraints on the order of matrices.*
It is clear that all matrices will be of order $m \times m$.

$$\begin{aligned} Y &= XPX + QX^2 + X \\ &= \left(XP + QX + I\right)X \end{aligned}$$

$$XPX + QXX + X$$

Product with Scalar and Quadratic Forms

Back to Definition 48 it is very important, sometimes, to make sure of conforming even for scalars; i.e., we write

$$y_{m\times 1}a_{1\times 1} \textbf{ NOT } ay.$$

This is because, sometimes, $a_{1\times 1}$ itself is a matrix multiplication that if dissembled it should conform with the remaining of equation

$$\begin{aligned} a_{1\times 1} &= x'_{1\times m}A_{m\times m}x_{m\times 1} \\ y_{n\times 1}a_{1\times 1} &= \underline{y_{n\times 1}x'_{1\times m}}A_{m\times m}x_{m\times 1} \\ a_{1\times 1}y_{n\times 1} &= x'_{1\times m}A_{m\times m}\underline{x_{m\times 1}y_{n\times 1}} \end{aligned} \qquad \text{(WRONG!)}$$

Example 77 *Expand and simplify* $y = \left(x-\mu\right)' \Sigma \left(x-\mu\right)$, *where* x *and* μ *are vectors and* Σ *is a symmetric matrix.*

$$
\begin{aligned}
y &= \left(x-\mu\right)' \Sigma \left(x-\mu\right) \\
&= \left(x'-\mu'\right) \Sigma \left(x-\mu\right) \\
&= x'\Sigma x - x'\Sigma\mu - \mu'\Sigma x + \mu'\Sigma\mu \\
&= x'\Sigma x - x'\Sigma\mu - \left(\mu'_{1\times p}\Sigma_{p\times p}x_{p\times 1}\right)' + \mu'\Sigma\mu \qquad (scalar' = scalar) \\
&= x'\Sigma x - x'\Sigma\mu - x'\Sigma\mu + \mu'\Sigma\mu \\
&= x'\Sigma x - 2x'\Sigma\mu + \mu'\Sigma\mu
\end{aligned}
$$

5.5 Projection

5.5.1 Projection on Single Vector

Suppose: $\mathbf{X}_{p\times 1} = X$. If $X' = (1,0,\ldots,0)$, we simply get the first component Y_1.

if $X' = (1/\sqrt{2}, 1/\sqrt{2})$, we project on the $\pi/4$ direction; this will be the value of the new vector in the direction of X.

$$\begin{aligned} X'Y &= \|X\|\,\|Y\|\cos(Y,X) \\ &= \|X\| \times \text{ Projected Length} \end{aligned}$$

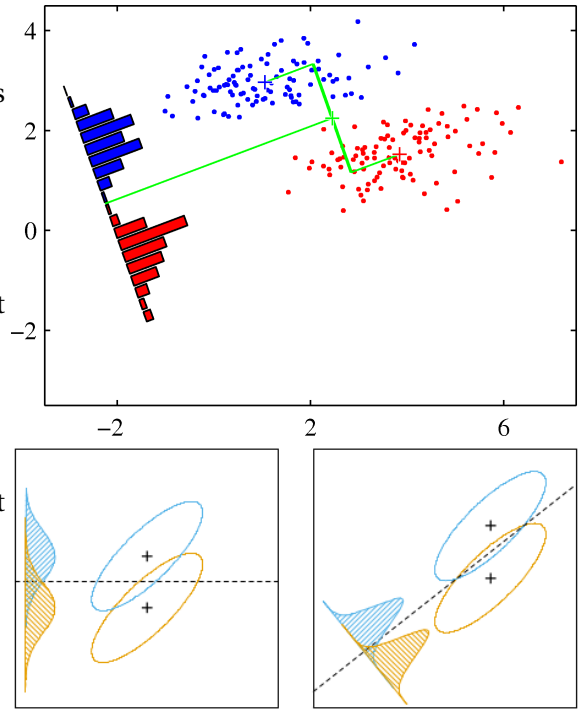
If we need the projected length only, then project on a unit vector $\frac{X'}{\|X\|}$, then

$$\frac{X'}{\|X\|}Y = \|Y\|\cos(Y,X).$$

Multiply this scalar in the direction of the projection to get the new component in the direction X

$$\widehat{Y} = \left(\frac{X}{\|X\|}\right)\left(\frac{X'}{\|X\|}Y\right) = \frac{XX'}{(X'X)}Y = \mathbf{P}_{p\times p}Y_{p\times 1},$$

where we call $\mathbf{P}$ the projection matrix of the direction X.



5.5.2 Projection on Set of Vectors (subspace)

When we express a vector in lower subspace of vectors, the columns of $\mathbf{X} = (X_1, \ldots, X_n)$:

$$\widehat{Y} = X_1\widehat{\beta}_1 + \ldots + X_n\widehat{\beta}_n$$

We need to minimize (this is **optimization**) the remaining error

$$e = Y - \mathbf{X}\widehat{\beta}.$$

After we differentiate w.r.t. $\widehat{\beta}$ and equate to zero we get:

$$e'\mathbf{X} = 0$$

, **perpendicular; wonderful!**. The solution will be

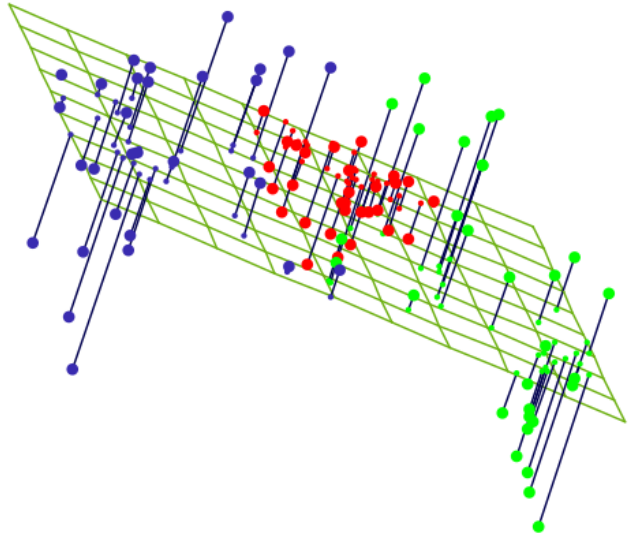
$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

Therefore, the projection $\widehat{Y}$ is

$$\widehat{Y} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

and the projection matrix $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$



Interestingly, when X_i and X_j are orthogonal:

$$\mathbf{P} = \frac{X_1X_1'}{X_1'X_1} + \ldots + \frac{X_nX_n'}{X_n'X_n} = \mathbf{P}_1 + \ldots + \mathbf{P}_n$$
$$\widehat{Y} = \mathbf{P}_1Y + \ldots + \mathbf{P}_nY.$$

Example 78 (Simulation using Python) *:*

This is a very good opportunity to emphasize the contemplation of Sec. 4.5

Chapter 6

A Snapshot of Multivariate Probability and Statistics

6.1 Random Vectors

A p-dimensional random vector X is $X = (X_1, \ldots, X_p)'$ has joint pdf

$$f_X = f_{X_1,\ldots,X_p}$$

Mean:

$$\begin{aligned}\mu &= \mathrm{E}X \\ &= (\mathrm{E}X_1, \ldots, \mathrm{E}X_p)'\end{aligned}$$

Covariance Matrix Σ $(= \mathrm{Cov}(X))$:

$$\begin{aligned}\Sigma &= \mathrm{E}(X-\mu)(X-\mu)' \\ &= \mathrm{E}\left[\begin{pmatrix} X_1-\mu_1 \\ \vdots \\ X_p-\mu_p \end{pmatrix}(X_1-\mu_1, \ldots, X_p-\mu_p)\right] \\ &= \mathrm{E}\begin{pmatrix} (X_1-\mu_1)^2 & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ \vdots & \ddots & \\ (X_p-\mu_p)(X_1-\mu_1) & & (X_p-\mu_p)^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \\ \sigma_{p1} & & \sigma_p^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \vdots & \ddots & \\ \rho_{1p}\sigma_1\sigma_p & & \sigma_p^2 \end{pmatrix}.\end{aligned}$$

Σ is symmetric; i.e., $\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \sigma_{j_i}$

6.2 Multinormal Distribution

X is said to have a multinormal distribution ($X \sim \mathcal{N}(\mu, \Sigma)$) if

$$f_X(x) = \frac{1}{((2\pi)^p |\Sigma|)^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)},$$
$$\mathrm{E}\, X = \mu, \qquad \mathrm{Cov}(X) = \Sigma.$$

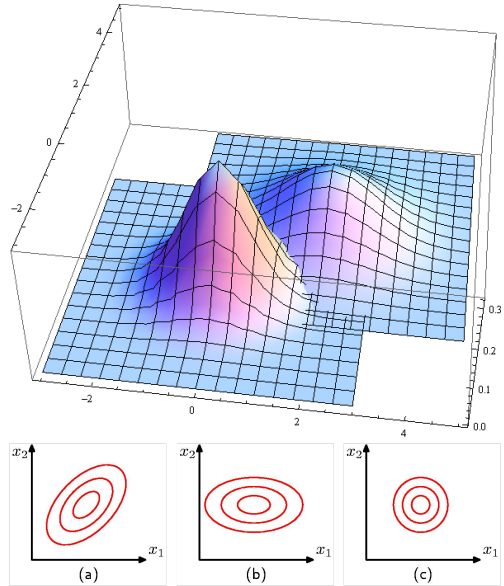
Remark: what is the case of $p = 1$?

You can generate a covariance matrix by choosing $\sigma_i^2, i = 1, \ldots, p$; then choose ρ_{ij}, where $-1 \le \rho_{ij} \le 1$. In the case of $p = 2$:
(a) $\rho > 0$
(b) $\rho = 0$, and $\sigma_2 < \sigma_1$.
(c) $\rho = 0$, and $\sigma_2 = \sigma_1$.

For diagonal Σ (uncorrelated components),

$$f_X(x) = \frac{1}{\prod_{i=1}^p \left(2\pi\sigma_i^2\right)^{1/2}} \exp\left[\sum_{i=1}^p -\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}\right]$$
$$= \prod_i \frac{1}{\left(2\pi\sigma_i^2\right)^{1/2}} \exp\left[-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}\right],$$

which is joint of p independent normals. This is not the case for other distributions.



Simulations, why, and example:
Revisit Example 78 for Σ building and data generation.

6.3 Estimation

Estimation of μ and Σ is nothing but estimation of their components; in vector form:

$$\widehat{\mu} \equiv \overline{X} = \frac{1}{n}\sum_i x_i = \frac{1}{n}\left[\begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix} + \ldots + \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}\right] = \begin{pmatrix} \frac{1}{n}\sum_i x_{i1} \\ \vdots \\ \frac{1}{n}\sum_i x_{ip} \end{pmatrix} = \begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_p \end{pmatrix}$$

$$\widehat{\Sigma} = \frac{1}{n-1}\sum_i \left(x_i - \overline{X}\right)\left(x_i - \overline{X}\right)' =$$

$$\begin{pmatrix} \frac{1}{n-1}\sum_i \left(x_{i1} - \overline{X}_1\right)^2 & \cdots & \frac{1}{n-1}\sum_i \left(x_{i1} - \overline{X}_1\right)\left(x_{ip} - \overline{X}_p\right) \\ \vdots & \ddots & \\ \frac{1}{n-1}\sum_i \left(x_{ip} - \overline{X}_p\right)\left(x_{i1} - \overline{X}_1\right) & & \frac{1}{n-1}\sum_i \left(x_{ip} - \overline{X}_p\right)^2 \end{pmatrix} = \begin{pmatrix} \widehat{\sigma_1^2} & \cdots & \widehat{\sigma_{1p}} \\ \vdots & \ddots & \\ \widehat{\sigma_{p1}} & & \widehat{\sigma_p^2} \end{pmatrix}$$

Usually, we put data as $\mathbf{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$, then:

$$\widehat{\mu}' = \frac{1}{n}\sum_i x_i'$$

$$\widehat{\Sigma} = \frac{1}{n-1}\left((x_1 - \widehat{\mu}) \ldots (x_n - \widehat{\mu})\right)\begin{pmatrix} (x_1 - \widehat{\mu})' \\ \vdots \\ (x_n - \widehat{\mu})' \end{pmatrix} = \frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c.$$

Example 79 (Running example to Data Science II) *:*

$$\widehat{\mu}' = \frac{1}{n}\sum_i x_i' = \frac{1}{n}\mathbf{1}'\mathbf{X}$$

$$= \frac{1}{n}\mathbf{1}' \begin{pmatrix} 0 & 2 \\ 2 & 6 \\ 2 & 7 \\ 2 & 5 \\ 4 & 9 \\ 4 & 8 \\ 4 & 7 \\ 6 & 10 \\ 6 & 11 \\ 6 & 9 \\ 8 & 15 \\ 8 & 13 \end{pmatrix}$$

$$= \begin{pmatrix} 4.3333 & 8.5 \end{pmatrix}$$

```
mu = np.mean(X,0)
```

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}\overline{X}' =$$

$$\begin{pmatrix} 0 & 2 \\ 2 & 6 \\ 2 & 7 \\ 2 & 5 \\ 4 & 9 \\ 4 & 8 \\ 4 & 7 \\ 6 & 10 \\ 6 & 11 \\ 6 & 9 \\ 8 & 15 \\ 8 & 13 \end{pmatrix} - \begin{pmatrix} 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \\ 4.333 & 8.5 \end{pmatrix} = \begin{pmatrix} -4.3 & -6.5 \\ -2.3 & -2.5 \\ -2.3 & -1.5 \\ -2.3 & -3.5 \\ -0.333 & 0.5 \\ -0.333 & -0.5 \\ -0.333 & -1.5 \\ 1.7 & 1.5 \\ 1.7 & 2.5 \\ 1.7 & 0.5 \\ 3.7 & 6.5 \\ 3.7 & 4.5 \end{pmatrix}$$

```
Xc = X-mu
```

$$\widehat{\Sigma} = \frac{1}{n-1}\left(\mathbf{X}_c'\mathbf{X}_c\right)$$

$$= \frac{1}{11}\begin{pmatrix} 70.667 & 94.0 \\ 94.0 & 137.0 \end{pmatrix} = \begin{pmatrix} 6.4242 & 8.5454 \\ 8.5454 & 12.4545 \end{pmatrix}$$

```
Sigma = 1/(n-1)*np.dot(Xc.T, Xc)
```

In one step:

```
Sigma = np.cov(X.T)
```

6.4 Transformation by Projection

Theorem 80 (Mean and Covariance Matrix after Transformation) *: For any random vector $X_{p\times 1}$, any matrix of m column vectors $A_{p\times m}$, as a set of m projections from $\mathbf{R}^p$ to $\mathbf{R}^m$ such that: $A'_{m\times p} = \begin{pmatrix} \alpha'_1 \\ \vdots \\ \alpha'_m \end{pmatrix}$,*

We have:

$$\mathrm{E}\, A'_{m\times p} X_{p\times 1} = A' \,\mathrm{E}\, X = A'\mu.$$
$$\mathrm{Cov}\left(A'_{m\times p} X_{p\times 1}\right) = A' \,\mathrm{Cov}\,(X)\, A = A'\Sigma A.$$

Theorem 81 (Normal Transformation) *: If $X \sim \mathcal{N}(\mu, \Sigma)$ then $A'X \sim \mathcal{N}(A'\mu, A'\Sigma A)$.*

Hints:

- The two theorems above are extension to the uni-variate case.
- Revisit Example 78

Chapter 7

A Snapshot of Optimization

7.1 Mathematical Optimization

Definition 82 *A mathematical optimization problem or just optimization problem, has the form (Boyd and Vandenberghe, 2004):*

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to:} \quad & f_i(x) \le 0, \qquad i = 1, \ldots, m \\ & h_i(x) = 0, \qquad i = 1, \ldots, p, \end{aligned}$$

$x = (x_1, \ldots, x_n) \in \mathbf{R}^n$, (*optimization variable*)

$f_0 : \mathbf{R}^n \mapsto \mathbf{R}$, (*objective (cost/utility) function*)

$f_i : \mathbf{R}^n \mapsto \mathbf{R}$, (*inequality constraints (functions)*)

$h_i : \mathbf{R}^n \mapsto \mathbf{R}$, (*equality constraints (functions)*)

$$\mathcal{D} : \bigcap_{i=1}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i \qquad \text{(\textit{feasible set})}$$

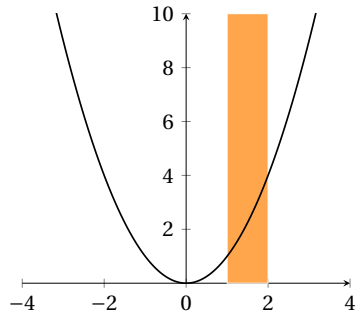
$$= \left\{ x \mid x \in \mathbf{R}^n \ \wedge \ f_i(x) \le 0 \ \wedge \ h_i(x) = 0 \right\}$$

$$x^* : \left\{ x \mid x \in \mathcal{D} \ \wedge \ f_0(x) \le f_0(z) \ \forall z \in \mathcal{D} \right\} \qquad \text{(\textit{solution})}$$

- minimize $f_0 \equiv$ maximize $-f_0$.
- $f_i \le 0 \ \equiv \ -f_i \ge 0$.
- 0s can be replaced of course by constants b_i, c_i
- unconstrained problem when $m = p = 0$.

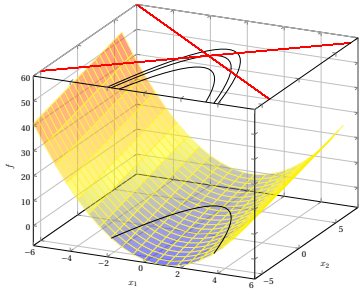
Example 83 *:*

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x^2 \\ \text{subject to:} \quad & x \le 2 \ \wedge \ x \ge 1. \end{aligned}$$



$x^* = 1$.

If the constraints are relaxed, then $x^* = 0$.



$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to:} \quad & f_i(x) \le 0, \qquad i = 1,\dots,m \\ & h_i(x) = 0, \qquad i = 1,\dots,p, \end{aligned}$$

$x = (x_1,\dots,x_n) \in \mathbf{R}^n$, (*optimization variable*)

$f_0 : \mathbf{R}^n \mapsto \mathbf{R}$, (*objective (cost/utility) function*)

$f_i : \mathbf{R}^n \mapsto \mathbf{R}$, (*inequality constraints (functions)*)

$h_i : \mathbf{R}^n \mapsto \mathbf{R}$, (*equality constraints (functions)*)

$$\mathcal{D} : \bigcap_{i=1}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i \qquad \text{(\textit{feasible set})}$$

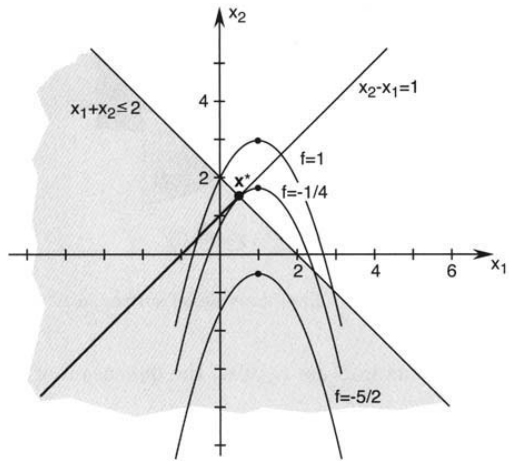
$$= \left\{ x \mid x \in \mathbf{R}^n \ \wedge\ f_i(x) \le 0 \ \wedge\ h_i(x) = 0 \right\}$$

$$x^* : \left\{ x \mid x \in \mathcal{D} \ \wedge\ f_0(x) \le f_0(z)\ \forall z \in \mathcal{D} \right\} \qquad \text{(\textit{solution})}$$

Example 84 *(Chong and Zak, 2013, Ex. 20.1, P. 454):*

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & (x_1-1)^2 + x_2 - 2 \\ \text{subject to:} \quad & x_2 - x_1 = 1 \\ & x_1 + x_2 \le 2. \end{aligned}$$

No global minimizer: $\partial z/\partial x_2 = 1 \neq 0$. *However,* $z|_{(x_2-x_1=1)} = (x_1-1)^2 + (x_1-1)$, *which attains a minima at* $x_1 = 1/2$.



$x* = (1/2, 3/2)'$. (Let's see animation)

7.1.1 Motivation and Applications

- *optimization problem* is an abstraction of how to make "best" possible choice of $x \in \mathbf{R}^n$.
- *constrains* represent trim requirements or specifications that limit the possible choices.
- *objective function* represents the *cost* to minimize or the *utility* to maximize for each x.

Examples:

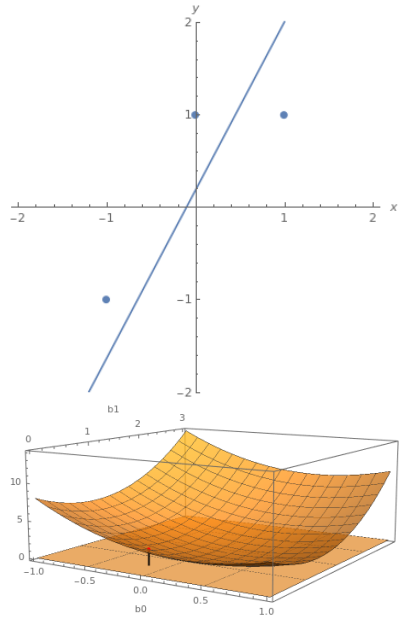
	Any problem	*Portfolio Optimization*	*Device Sizing*	*Data Science*
$x \in \mathbf{R}^n$	choice made	investment in capitals	dimensions	parameters
$f_i,\ h_i$	firm requirements /conditions	overall budget	engineering constraints	regularizer
f_0	cost (or utility)	overall risk	power consumption	error

- Amazing variety of practical problems. In particular, data science: two sub-fields: construction and assessment.
- The construction of: Least Mean Square (LMS), Logistic Regression (LR), Support Vector Machines (SVM), Neural Networks(NN), Deep Neural Networks (DNN), etc.
- Many techniques are for solving the optimization problem:
  - Closed form solutions: convex optimization problems
  - Numerical solutions: Newton's methods, Gradient methods, Gradient descent, etc.
  - "Intelligent" methods: particle swarm optimization, genetic algorithms, etc.

Example 85 (Machine Learning: construction) *:*
Let's suppose that the best regression function is $Y = \beta_0 + \beta_1 X$*, then for the training dataset* (x_i, y_i) *we need to minimize the MSE.*

- Half of ML field is construction: NN, SVM, etc.
- In DNN it is an optimization problem of millions of parameters.
- Let's see animation.
- Where are Probability, Statistics, and Linear Algebra here? Let's re-visit the chart.
- Is the optimization problem solvable:
  - closed form? (LSM)
  - numerically and guaranteed? (convex and linear)
  - numerically but not guaranteed? (non-convex):
    - numerical algorithms, e.g., GD,
    - local optimization,
    - heuristics, swarm, and genetics,
    - brute-force with exhaustive search

$$\underset{\beta_o,\beta_1}{\text{minimize}} \sum_i (\beta_o + \beta_1 x_i - y_i)^2$$



7.1.2 Solving Optimization Problems

- A *solution method* for a class of optimization problems is an algorithm that computes a solution.
- Even when the *objective function* and constraints are smooth, e.g., polynomials, the solution is very difficult.
- There are three classes where solutions exist, theory is very well developed, and amazingly found in many practical problems:

$$\text{Linear} \subset \text{Quadratic} \subset \text{Convex} \subset \text{Non-linear (not linear and not known to be convex!)}$$

- For the first three classes, the problem can be solved very reliably in hundreds or thousands of variables!

7.2 Least-Squares and Linear Programming

7.2.1 Least-Squares Problems

A *least-squares* problem is an optimization problem with no constraints (i.e., $m = p = 0$), and an objective in the form:

$$\underset{x}{\text{minimize}} f_0(x) = \sum_{i=1}^{k} (a_i' x - b_i)^2 = \|A_{k\times n} x_{n\times 1} - b_{k\times 1}\|^2.$$

The solution is given in **closed form** by:

$$x = (A'A)^{-1} A'b$$

- Good algorithms in many SC SW exist; it is a very mature technology.
- Solution time is $O(n^2 k)$.
- Easily solvable even for hundreds or thousands of variables.
- More on that in the Linear Algebra course.
- Many other problems reduce to typical LS problem:
  - Weighted LS (to emphasize some observations)

$$\underset{x}{\text{minimize}} f_0(x) = \sum_{i=1}^{k} w_i (a_i' x - b_i)^2.$$

  - Regularization (to penalize for over-fitting)

$$\underset{x}{\text{minimize}} f_0(x) = \sum_{i=1}^{k} (a_i' x - b_i)^2 + \rho \sum_{j=1}^{n} x_j^2.$$

7.2.2 Linear Programming

A *linear programming* problem is an optimization problem with objective and all constraint functions are linear:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) = C'x \\ \text{subject to:} \quad & a_i'x \le b_i, && i = 1,\dots,m \\ & h_i'x = g_i, && i = 1,\dots,p, \end{aligned}$$

- **No** closed form solution as opposed to LS.
- Very robust, reliable, and effective set of methods for numerical solution; e.g., Dantzig's simplex, and interior point.
- Complexity is $\simeq O(n^2m)$.
- Similar to LS, we can solve a problem of thousands of variables.
- Example is *Chebyshev minimization* problem:

$$\underset{x}{\text{minimize}} f_0(x) = \max_{i=1,\dots,k} |a_i'x - b_i|,$$

- The objective is different from the LS: minimize the maximum error. **Ex:**
- After some tricks, requiring familiarity with optimization, it is equivalent to a LP:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & t \\ \text{subject to:} \quad & a_i'x - t \le b_i, && i = 1,\dots,k \\ & -a_i'x - t \le -b_i, && i = 1,\dots,k \end{aligned}$$

7.3 Convex Optimization

A *convex optimization* problem is an optimization problem with objective and all constraint function are convex:

$$
\begin{array}{lll}
\underset{x}{\text{minimize}} & f_0(x) & \\
\text{subject to:} & f_i(x) \le 0, & i = 1,\dots,m \\
& h_i(x) = 0, & i = 1,\dots,p, \\
& f_i(\alpha x + \beta y) \le \alpha f_i(x) + \beta f_i(y), \qquad \alpha+\beta = 1, & 0 \le \alpha,\ 0 \le \beta, \qquad 0 \le i \le m \\
& h_i(x) = a_i' x + b_i & 0 \le i \le p
\end{array}
$$

- The LP and LS are special cases; however, only LS has closed-form solution.
- Very robust, reliable, and effective set of methods, including *interior point methods*.
- Complexity is almost: $O\big(\max(n^3, n^2 m, F)\big)$, where F is the cost of evaluating 1st and 2nd derivatives of f_i and h_i.
- Similar to LS and LP, we can solve a problem of thousands of variables.
- However, it is not as very mature technology as the LP and LS yet.
- There are many practical problems that can be re-formulated as convex problem **BUT** requires mathematical skills; but once done the problem is solved. **Hint:** realizing that the problem is convex requires more mathematical maturity than those required for LP and LS.

7.4 Nonlinear Optimization

A *non-linear optimization* problem is an optimization problem with objective and constraint functions are non-linear **BUT** not known to be convex (**so far**). Even simple-looking problems in 10 variables can be extremely challenging. Several approaches for solutions:

Local Optimization : starting at initial point in space, using differentiablity, then navigate

- does not guarantee global optimal.
- affected heavily by initial point.
- depends heavily on numerical algorithm and their parameters.
- More art than technology.
- In contrast to convex optimization, where a lot of art and mathematical skills are required to formulate the problem as convex; then numerical solution is straightforward.

Global Optimization : the true global solution is found; the compromise is complexity.

- The complexity goes exponential with dimensions.
- Sometimes it is worth it when: the cost is huge, not in real time, and dimensionality is low.

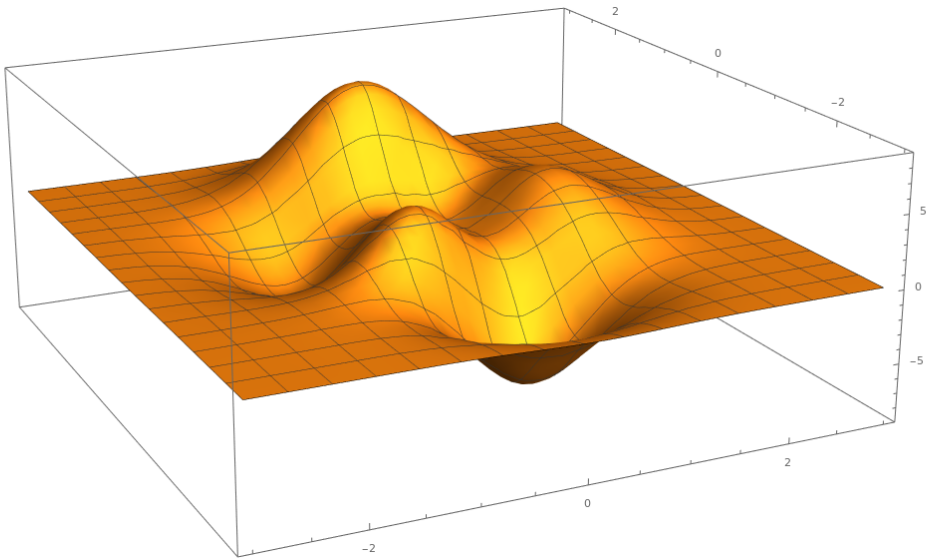
Role of Convex Optimization :

- Approximate the non-linear function to a convex one, finding the exact solution, then using it as a starting point for the original problem. (Also does not guarantee optimality)
- Setting bounds on the global solution.

Evolutionary Computations : Genetic Algorithm (GA), Simulated Annealing (SA), Particle Swarm Optimization (PSO), etc.

Example 86 (Nonlinear Objective Function) *: (Chong and Zak, 2013, Ex. 14.3, P.290)*

$$f(x,y) = 3(1-x)^2 e^{-x^2-(y+1)^2} - 10e^{-x^2-y^2}\left(-x^3 + \frac{x}{5} - y^5\right) - \frac{1}{3}e^{-(x+1)^2-y^2}$$



Bibliography

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press.

Cherkassky, V. S. and Mulier, F. (1998), *Learning from data : concepts, theory, and methods*, New York: Wiley.

Chong, E. K. and Zak, Stanislaw, H. (2013), *An Introduction to Optimization*, Wiley-Interscience, 4th ed.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001), *Pattern classification*, New York: Wiley, 2nd ed.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, New York: Springer, 2nd ed.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (2000), *Understanding robust and exploratory data analysis*, New York: Wiley, wiley clas ed.

Rice, J. A. (2007), *Mathematical statistics and data analysis*, Belmont, CA: Thomson/Brooks/Cole, 3rd ed.