Mathematics of Big Data, I

Lecture 3: Review Probability, GLMs (conti), Schur Complement, Multivariate Gaussian Distribution

Weiqing Gu
Professor of Mathematics
Director of the Mathematics Clinic

Harvey Mudd College
Summer 2017

Today

- **Review Probability**
  - View Probability functions as special kind of functions
    - Binomial
    - Multinomial
    - Poisson
    - Beta distribution
  - Key characteristics
  - Conditional probability
- **Generalized Linear Model (GLMs) (continued)**
- **Schur's Complement**
- **Conditional Normal Distributions**
    - Review: Single variable normal distribution (i.e. Gaussian distribution) and Multivariate Gaussian Distribution

A probability function is a special function which must satisfy:

$$0 \leq P(X) \leq 1$$

$$\sum P(X) = 1$$

A Big Picture of Probability Theory

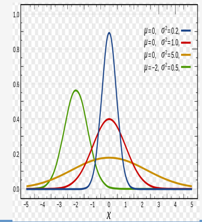
$0 \le P(X) \le 1$

$\sum P(X) = 1$

Key Characteristics:

Single rv	Muliti-rv
E(X) & Condit'l Expec'n	Cov (X, Y)
Variance/Stan. Devi.	Corrl(X, Y)
Moments	Cov. Matrix
Skewness etc.	Corrl Matrix

Probability Distributions (Discrete & Continuous) and their Geometric Meanings



Other known distrib'ns

- Bernoulli
- Beta $\theta \sim \text{Beta}(a, b)$
- Chi-square
- Poisson
- Student's t
- Uniform

Gaussian Distrib.

Discrete distrib'n → Taking limit → Conti. distrib'n

Probability Rules for Events:

Product rule/iid

Joint probability

Conditional Independence

	Discrete	Continuous
Single-rv	**Binomial** $\binom{n}{k} p^k (1-p)^{n-k}$	**Gaussian/Normal** $\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Multi-rv	**Multinomial** $\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$	**Multivari-Gaussian** $(2\pi)^{-\frac{1}{2}k} \lvert\Sigma\rvert^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$

Besides **pmf/pdf**, + 3 key fcns:

- **cdf** (cumulative distri. fcn)
- **cf** (characteristic fcn $E(e^{itX})$)
- **mgf** (moment generating fcn)

$m_X(t) = E(e^{tX})$

- **Central Limit Theorem**

Condi. Prob & **Bayesian Rules**

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

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

$$= \frac{p(X=x)p(Y=y|X=x)}{\sum_{x'} p(X=x')p(Y=y|X=x')}$$

Other Key Tech: Making connection to derivative/Jacobian/integrations.

Key: View everything as functions. P eats an observation x of a random variable X and spits out a value P(X=x) in [0,1], & the sum of all p(x) is 1.

- *X is a random variable. P(X=x) = p(x).*

Like the variables in calculus, we can add, subtract, make linear combinations; or make new functions f(x), also can take derivatives/integrations.

X→ f(X). For e.g.s

$f(X) = \sum a_i X_i$

$f(\mathbf{X}) = A\mathbf{X} + b$

$f(X) = X^n$

f(X) = Taylor exp.

what is *E*(f(X))?

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{cov}[\mathbf{y}] = \text{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) \left| \det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) \right| = p_x(\mathbf{x}) |\det \mathbf{J}_{\mathbf{y}\to\mathbf{x}}|$$

$$\mathbf{y} = f(\mathbf{x})$$

Two different ways to generalize Binomial distribution

- From Binomial distribution to Poisson distribution
- From Binomial distribution to Multinomial Distribution

- **Recall: What are Multinomial distributions?**
- **For example:** If a 6 sided die has
  - 3 faces painted red
  - 2 faces painted white
  - 1 faces painted blue

  And rolled 100 times.

  Find P(60 red, 30 white, and 10 blue).

  Work out details with the students on the board.

  Generally an experiment with m outcomes with respective probabilities $p_1, p_2, \ldots, p_m$ is performed n times independently.

  Let x_i = # of times outcome i appears, i=1,2,…,m

  Then $P(x_1=k_1, x_2=k_2, \ldots, x_m = k_m) = ?$

Claim: Multinomial distributions as exponential family distributi

Claim: Multinomial distributions as exponential family distribution.

- *Work out details with the students on the board.*

correlation coefficient & correlation matrix

- The (Pearson) **correlation coefficient** between two rvs X and Y is defined as

$$\operatorname{corr}[X,Y] \triangleq \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}$$

- If X and Y are indep., then cov [X, Y] = 0; say X and Y are uncorrelated.

- A **correlation matrix** of a random vector has the form:

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1,X_1] & \operatorname{corr}[X_1,X_2] & \cdots & \operatorname{corr}[X_1,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d,X_1] & \operatorname{corr}[X_d,X_2] & \cdots & \operatorname{corr}[X_d,X_d] \end{pmatrix}$$

Exercise: show that $-1 \leq \operatorname{corr}[X,Y] \leq 1$ and

Show that $\operatorname{corr}[X,Y] = 1$ iff $Y = aX + b$ for some parameters a and b.

Example of Correlation Coefficients

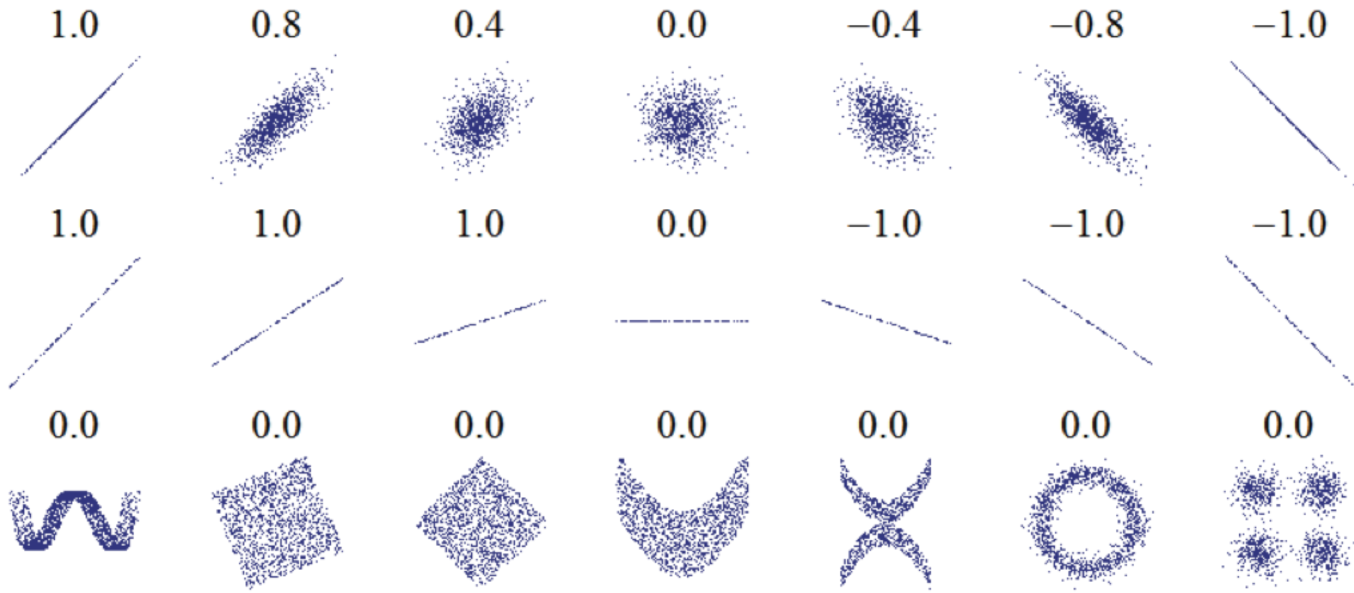


Figure 2.12 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

Conditional Probability

The **conditional probability** of event A, given that event B is true:

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

Bayes rule:

$$p(X=x|Y=y) = \frac{p(X=x,Y=y)}{p(Y=y)} = \frac{p(X=x)p(Y=y|X=x)}{\sum_{x'} p(X=x')p(Y=y|X=x')}$$

Recall: Probability of an Event

- p(A) denotes the probability that the event A is true.
- For example:
- A = a logical expression "it will rain tomorrow"

We require that $0 \leq p(A) \leq 1$.

p(A) = 0 means the event definitely will not happen

p(A) = 1 means the event definitely will happen

$p(\overline{A})$ denotes the probability of the event not A

$$p(\overline{A}) = 1 - p(A)$$

We also write:

A=1 to mean the event A is true.

A=0 to mean the event A is false.

Recall: Fundamental Rules

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned}$$

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

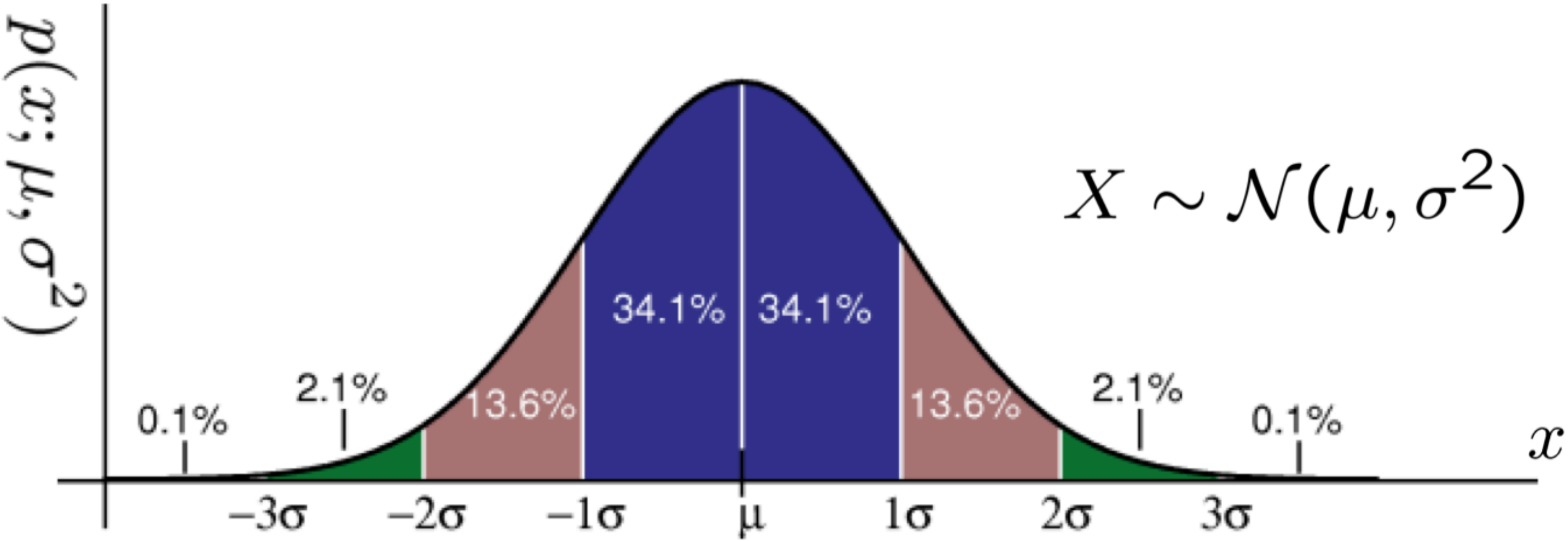
$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3) \ldots p(X_D|X_{1:D-1})$$

Changing gear:

Recall: Gaussian with one variable

(called *Univariate Gaussian*)

Gaussian distribution with mean μ, and standard deviation σ.



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

When μ = 0 and σ = 1, it is call the standard normal distribution.

Different ways to find expected values

$$E[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

Where f(x) is the probability density function of X.

Example: Let f(x) be the density of the standard normal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \mathrm{d}x$$

Method 1: Since is $xe^{-x^2/2}$ an odd function and the limits of the integral are symmetric, so we get E[X] =0.

Method 2: Directly integrate.

Method 3: Using the moment generating function.

Method 2

$$E[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{\frac{-x^2}{2}} \mathrm{d}x$$

$$= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} d(-\frac{x^2}{2})$$

$$= -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Big|_{-\infty}^{\infty}$$

$$= 0$$

Method 3

- The moment generating function is defined as

$$\phi(t) = E[e^{tX}].$$

$$\phi(t) = C\int_{\mathbb{R}} e^{tx}e^{-x^2/2}dx = C\int_{\mathbb{R}} e^{-x^2/2+tx}dx = e^{t^2/2}C\int_{\mathbb{R}} e^{-(x-t)^2/2}dx.$$

$$t^2/2 - (x-t)^2/2 = t^2/2 + (-x^2/2 + tx - t^2/2) = -x^2/2 + tx$$

1

$$\phi(t) = e^{t^2/2} = 1 + (t^2/2) + \frac{1}{2}\left(t^2/2\right)^2 + \cdots + \frac{1}{k!}\left(t^2/2\right)^k + \cdots.$$

$$E[e^{tX}] = E\left[1 + tX + \frac{1}{2}(tX)^2 + \cdots + \frac{1}{n!}(tX)^n + \cdots\right]$$

$$= 1 + E[X]t + \frac{1}{2}E[X^2]t^2 + \cdots + \frac{1}{n!}E[X^n]t^n + \cdots.$$

E[x] = 0

When k =1,
$E[x^2]$ =1.
Variance = 1.

Compare:

$$\frac{1}{(2k)!}E[X^{2k}]t^{2k} = \frac{1}{k!}(t^2/2)^k = \frac{1}{2^k k!}t^{2k},$$

$$E[X^{2k}] = \frac{(2k)!}{2^k k!},\quad k = 0, 1, 2, \ldots$$

Properties of Gaussians

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

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

- Integration of the densities equals to 1.

$$\int_{-\infty}^{\infty} p(x; \mu, \sigma^2) dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx = 1$$

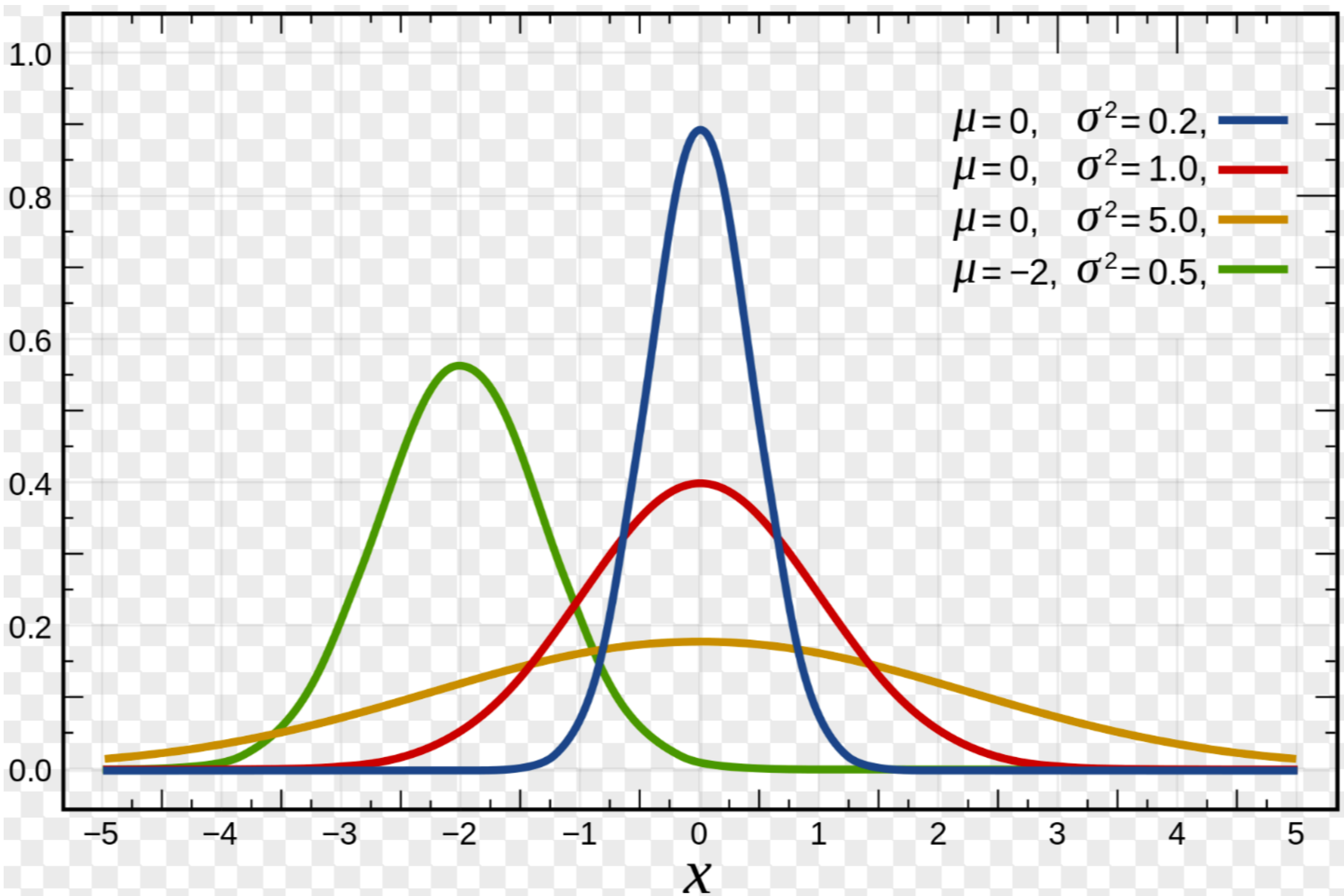
- Mean:

$$\begin{aligned} \mathrm{E}_X[X] &= \int_{-\infty}^{\infty} x p(x; \mu, \sigma^2) dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx \\ &= \mu \end{aligned}$$

- Variance:

$$\begin{aligned} \mathrm{E}_X[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x; \mu, \sigma^2) dx \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx \\ &= \sigma^2 \end{aligned}$$

In general, do translation and scale; i.e. change of variables when try to find those key characteristic values



Covariance, and Covariance Matrix

- The **covariance** between two rv's X and Y measures the degree to which X and Y are (linearly) related; defined as

$$\operatorname{cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Exercise

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

If **x** is a d-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

$$\operatorname{cov}[\mathbf{x}] \triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right]$$

$$= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\ \operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix}$$

Ofen denoted by Σ

correlation coefficient & correlation matrix

- The (Pearson) **correlation coefficient** between two rvs X and Y is defined as

$$\text{corr}[X,Y] \triangleq \frac{\text{cov}[X,Y]}{\sqrt{\text{var}[X]\,\text{var}[Y]}}$$

- If X and Y are indep., then cov [X, Y] = 0; say X and Y are uncorrelated.

- A **correlation matrix** of a random vector has the form:

$$\mathbf{R} = \begin{pmatrix} \text{corr}[X_1,X_1] & \text{corr}[X_1,X_2] & \cdots & \text{corr}[X_1,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}[X_d,X_1] & \text{corr}[X_d,X_2] & \cdots & \text{corr}[X_d,X_d] \end{pmatrix}$$

Exercise: show that $-1 \leq \text{corr}[X,Y] \leq 1$ and
Show that $\text{corr}[X,Y] = 1$ iff $Y = aX + b$ for some parameters a and b.

Example of Correlation Coefficients

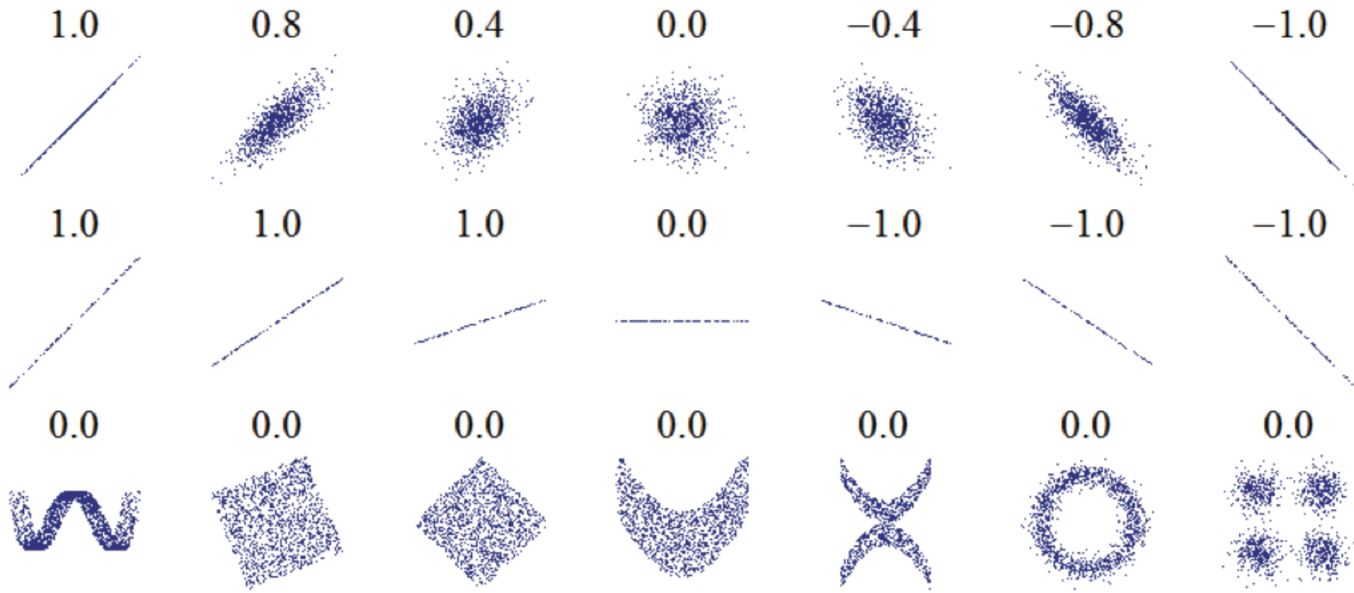


Figure 2.12 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

The multivariate Gaussian (distribution) or multivariate normal (MVN)

(The most widely used joint probability density function for continuous variables)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

determinant

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$

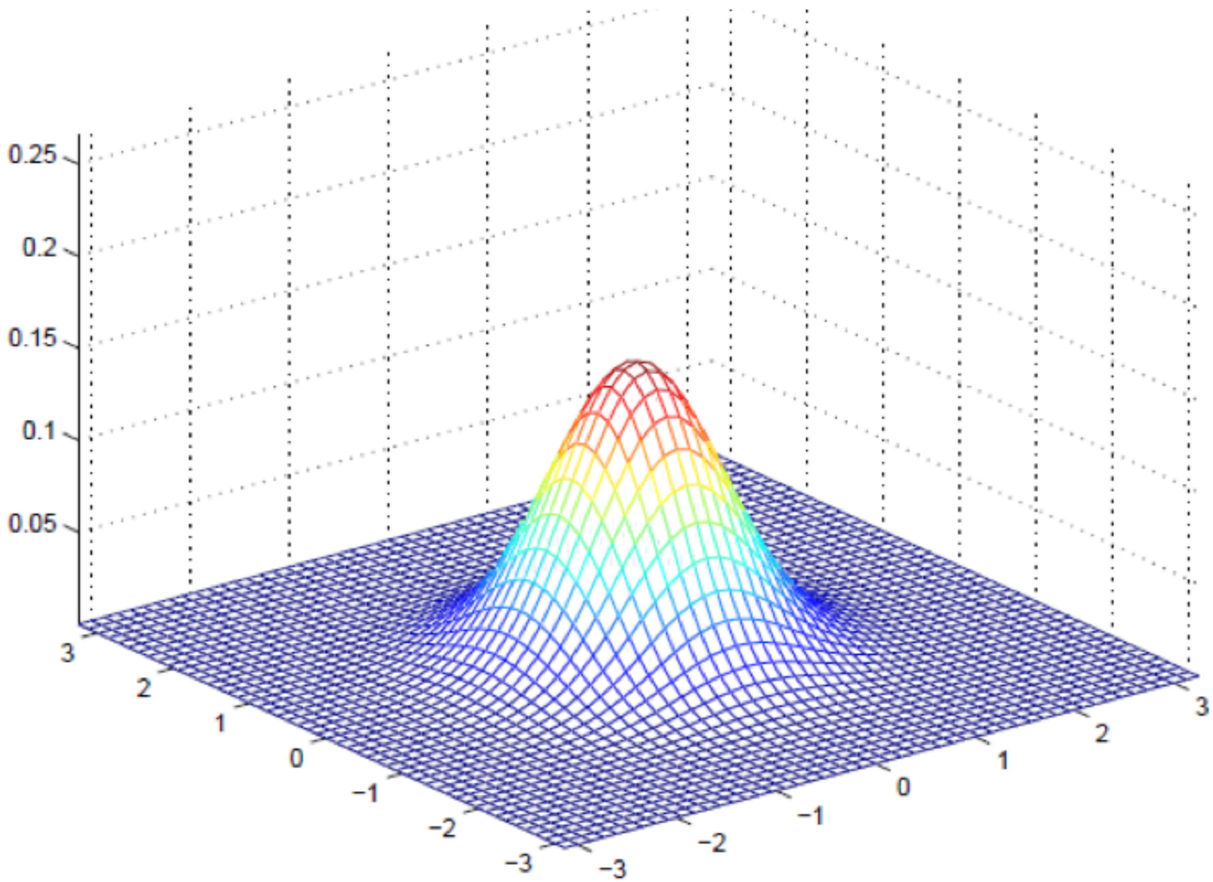
Note: **the precision matrix or concentration matrix is just**

the inverse covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$

A **spherical or isotropic covariance** $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_D$, has one free parameter.

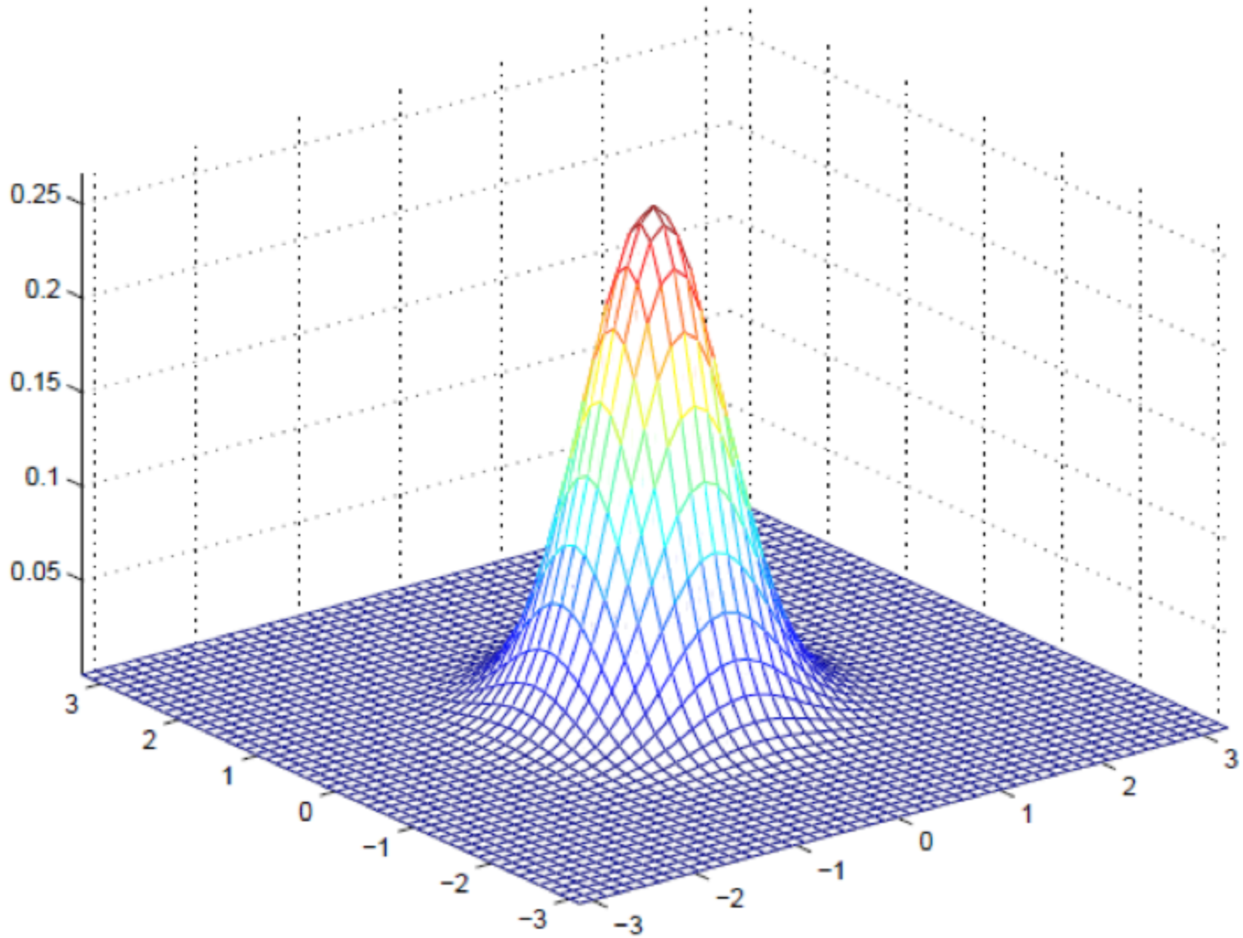
$$\mu = [0; 0]$$

$$\Sigma = [1\ 0\ ;\ 0\ 1]$$



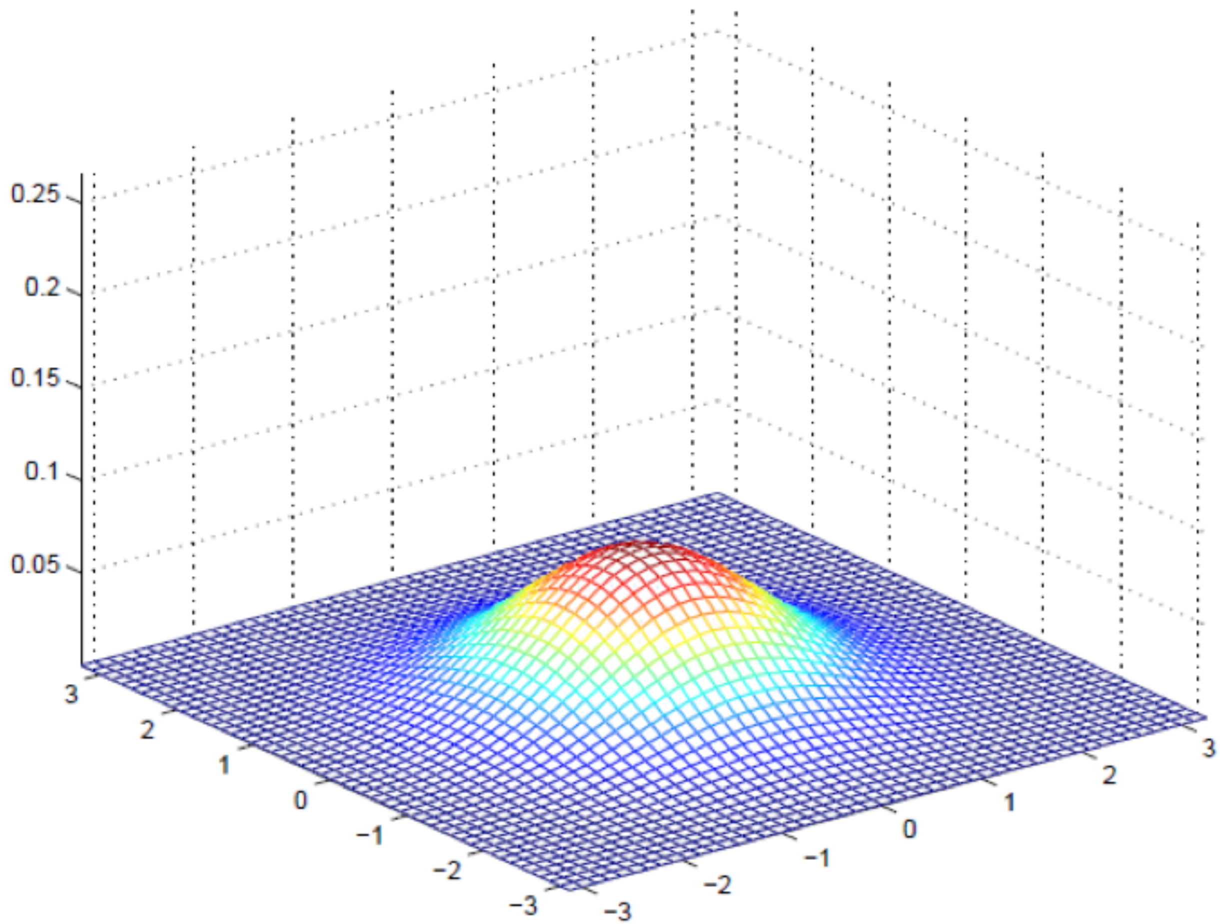
$$\mu = [0;\ 0]$$

$$\Sigma = [.6\ 0\ ;\ 0\ .6]$$



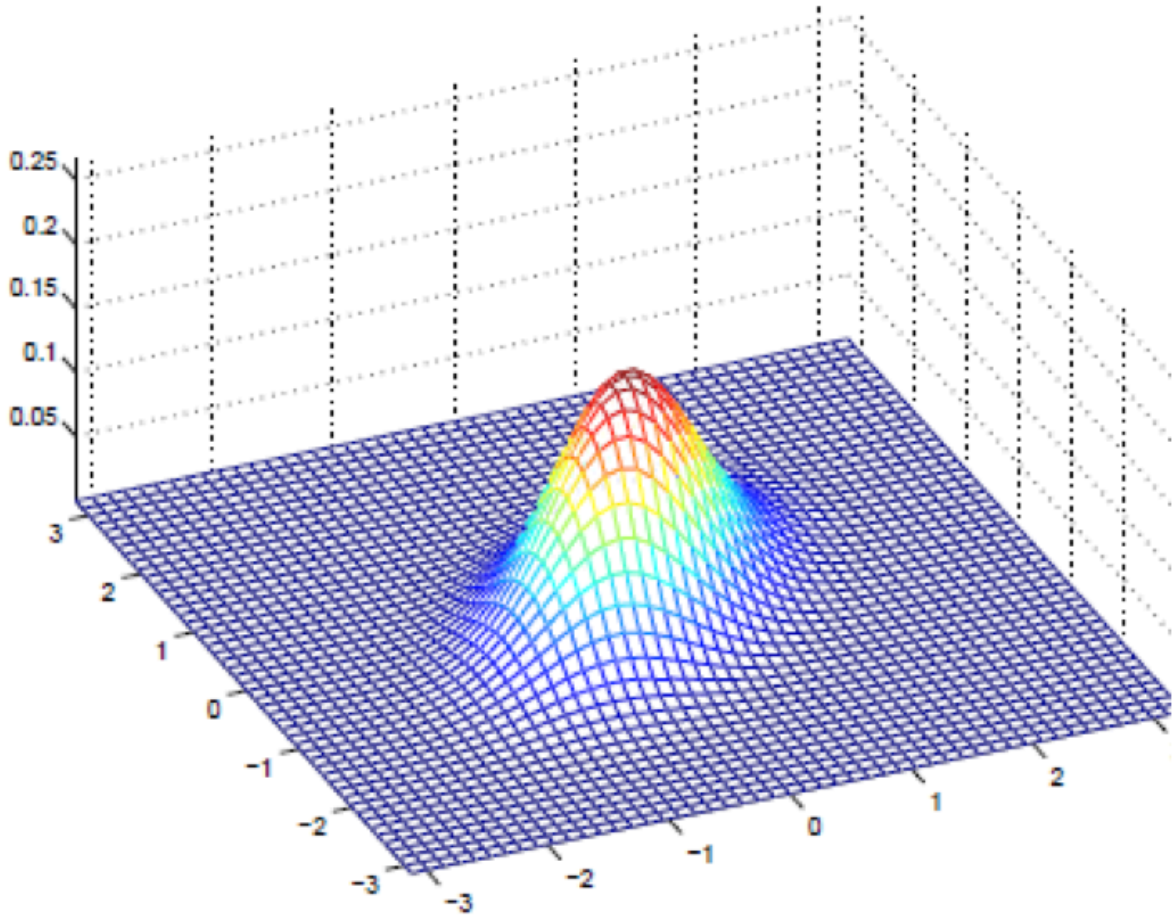
$$\mu = [0; 0]$$

$$\Sigma = [2\ 0\ ;\ 0\ 2]$$



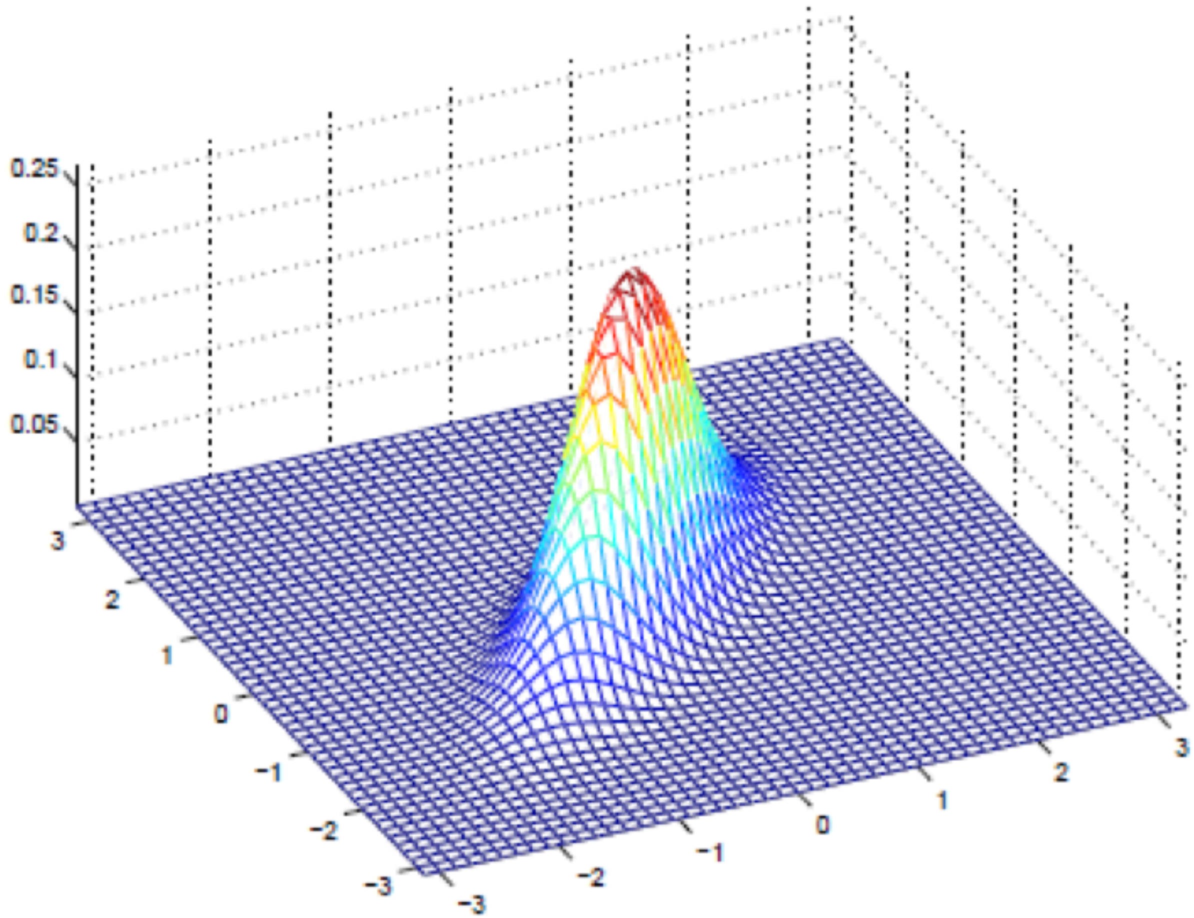
μ = [0; 0]

Σ = [1 0.5; 0.5 1]



$\mu = [0; 0]$

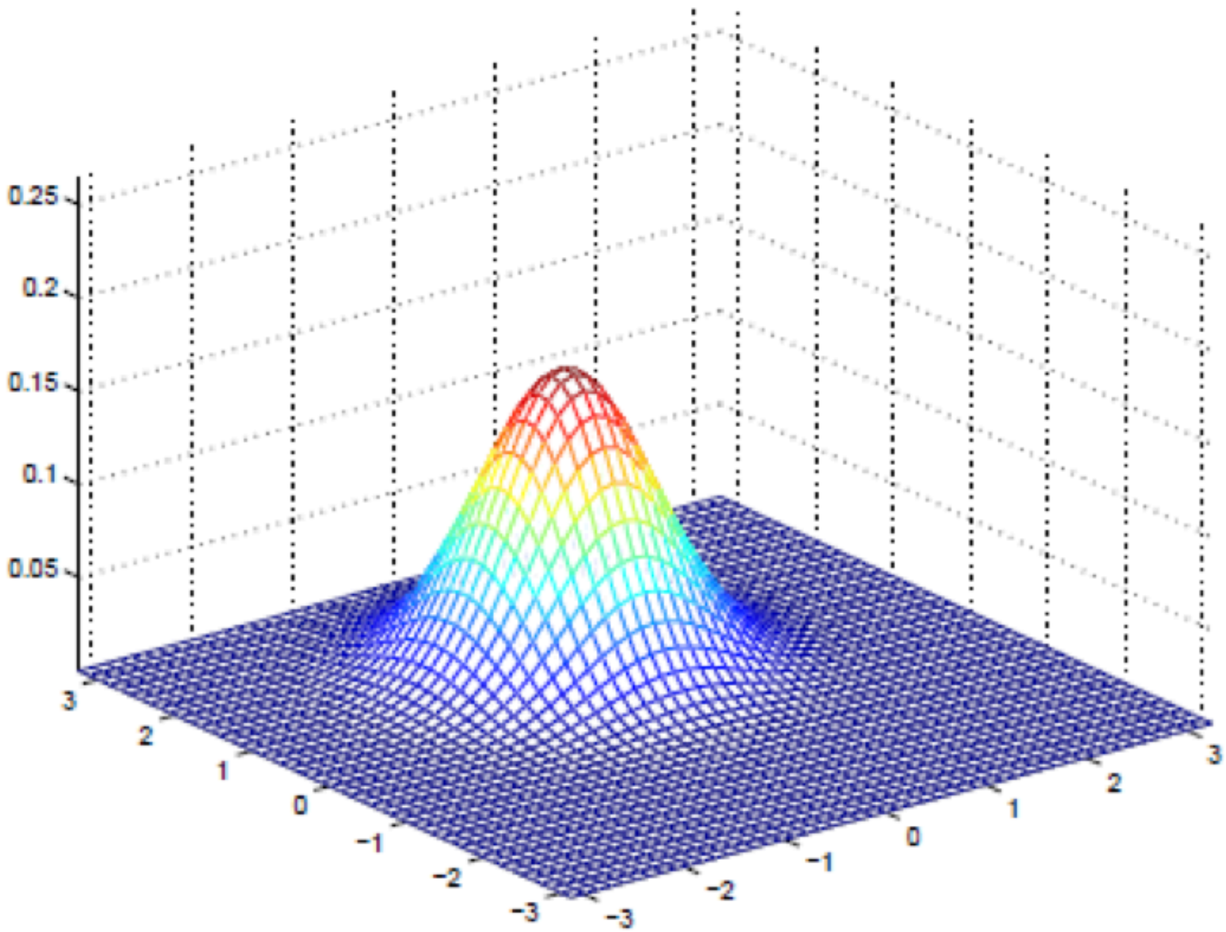
$\Sigma = [1 \;\; 0.8; 0.8 \;\; 1]$



Now let's visualize as μ changes

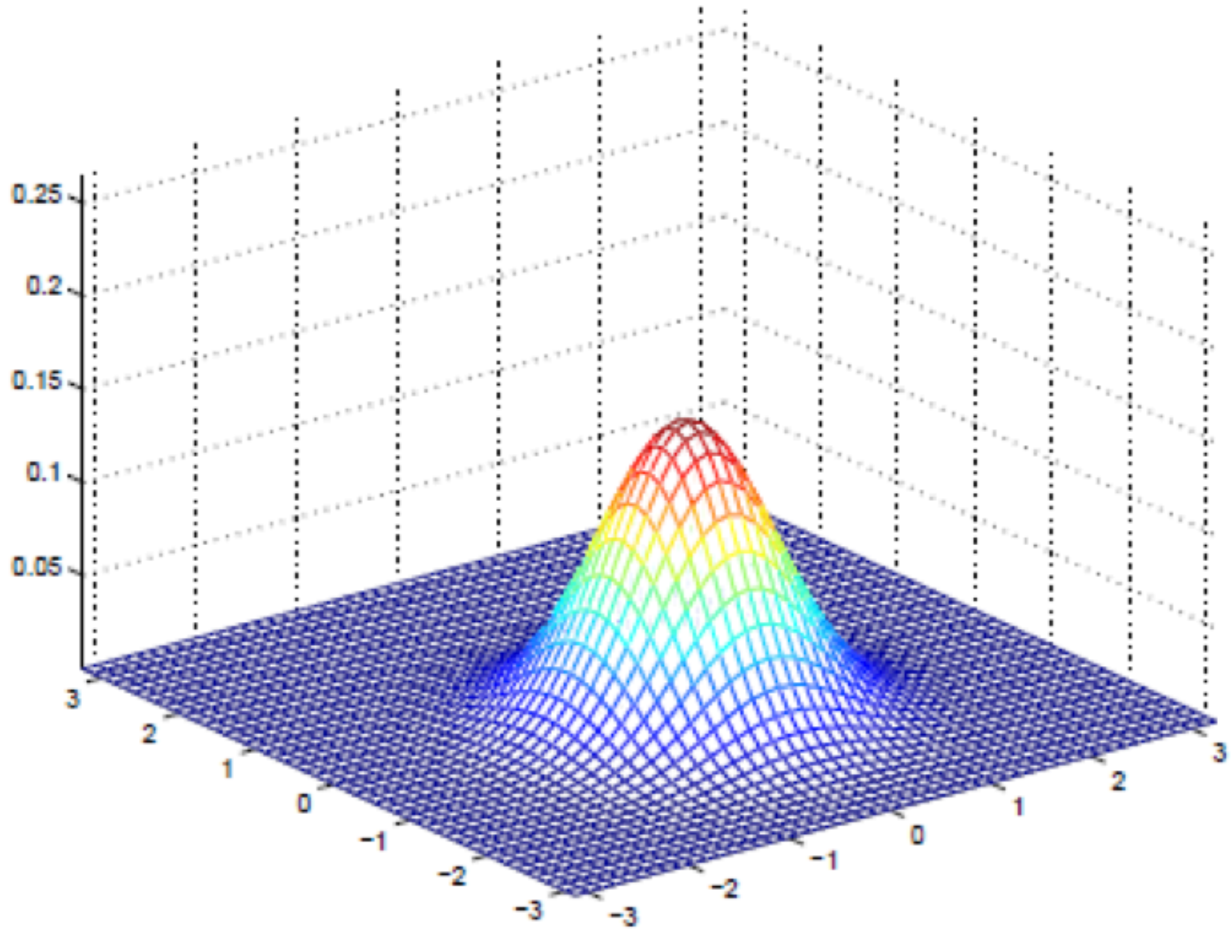
$$\mu = [1;\ 0]$$

$$\Sigma = [1\ \ 0;\ 0\ \ 1]$$



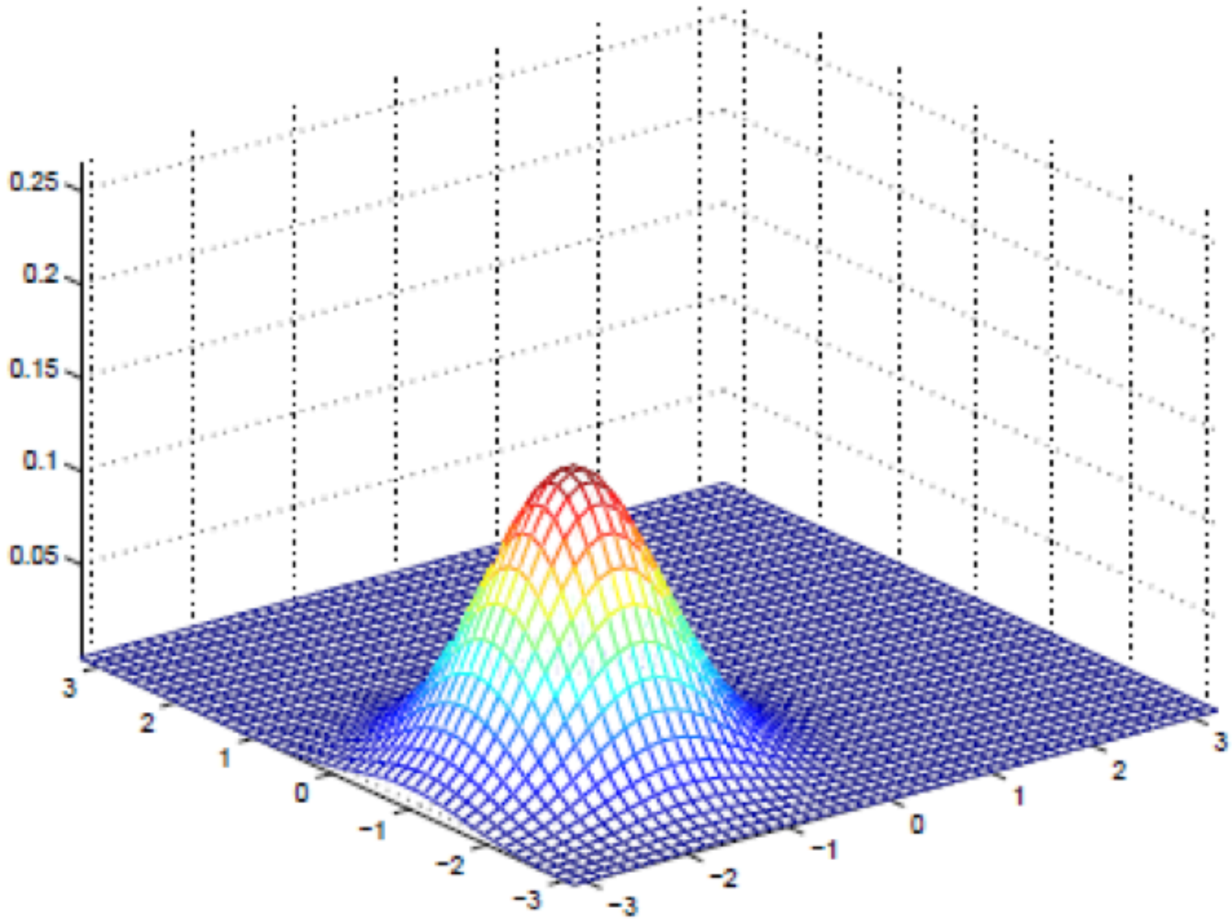
$\mu = [-.5; 0]$

$\Sigma = [1 \;\; 0; 0 \;\; 1]$



$\mu = [-1; -1.5]$

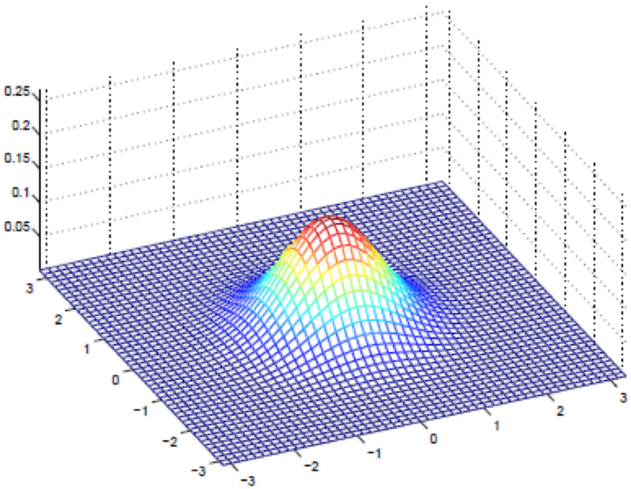
$\Sigma = [1 \;\; 0; 0 \;\; 1]$

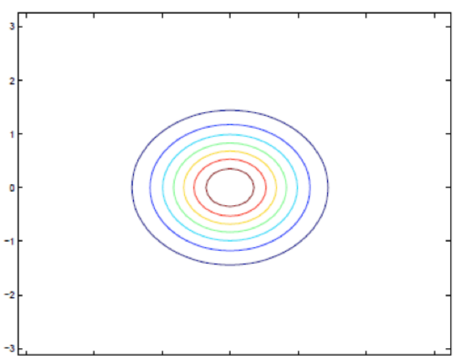


Level sets visualization

$\mu = [0; 0]$

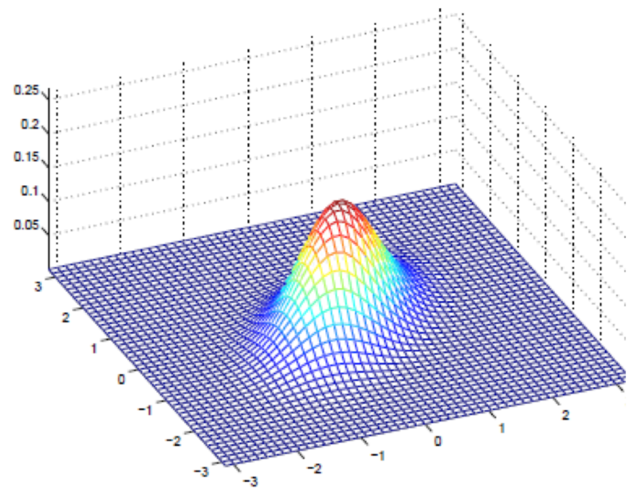
$\Sigma = [1 \;\; 0; 0 \;\; 1]$

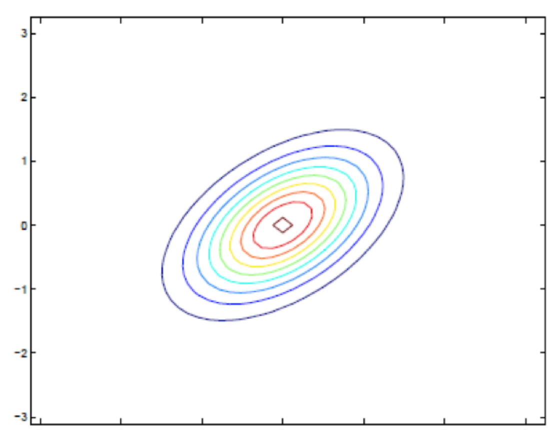




$\mu = [0; 0]$

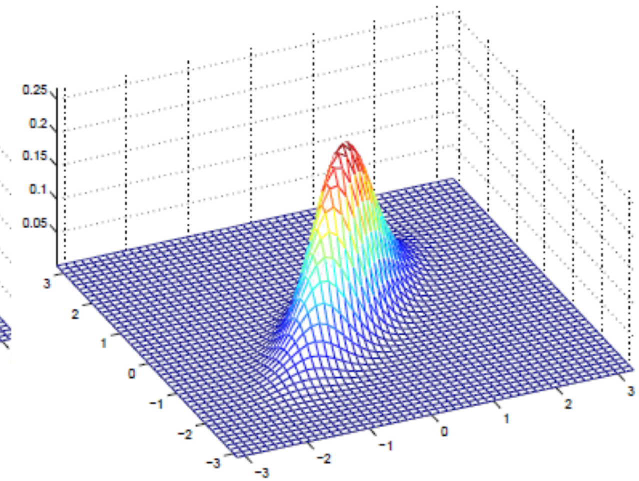
$\Sigma = [1 \;\; 0.5; 0.5 \;\; 1]$

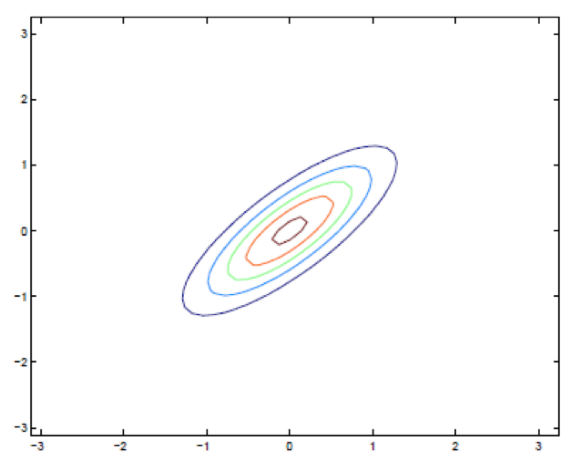




$\mu = [0; 0]$

$\Sigma = [1 \;\; 0.8; 0.8 \;\; 1]$



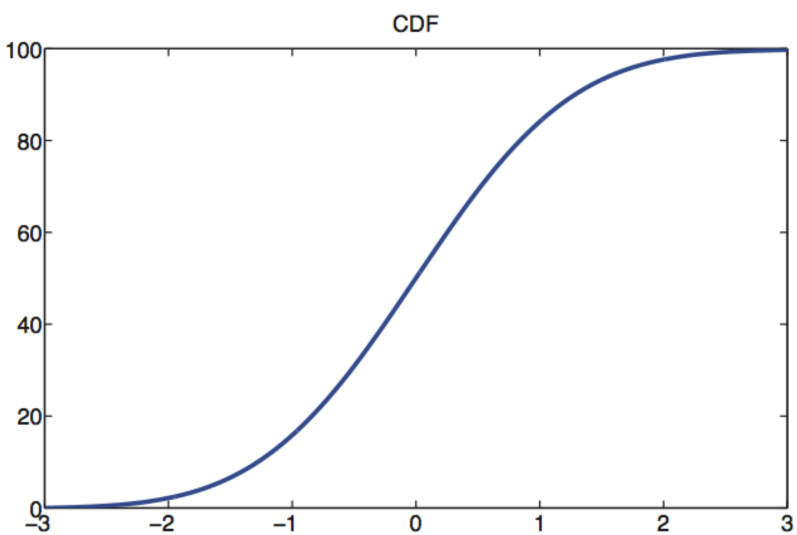


The cumulative distribution function (cdf)

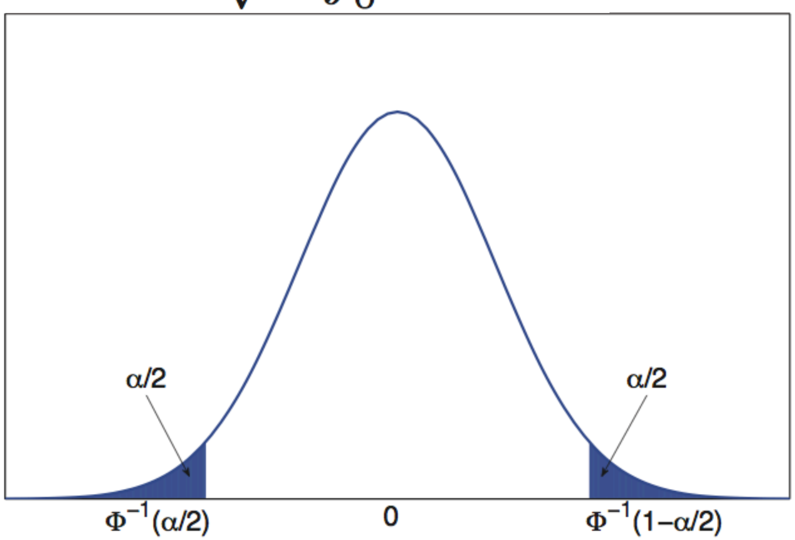
- For Gaussian distribution: $\Phi(x;\mu,\sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu,\sigma^2)dz$
- This integral has no closed form expression, but is built in to most software packages.

$$\Phi(x;\mu,\sigma) = \frac{1}{2}[1 + \text{erf}(z/\sqrt{2})]$$ where $z = (x-\mu)/\sigma$ and

$$\text{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$



(a) Plot of the cdf for the standard normal, $\mathcal{N}(0,1)$.



(b) Corresponding pdf.

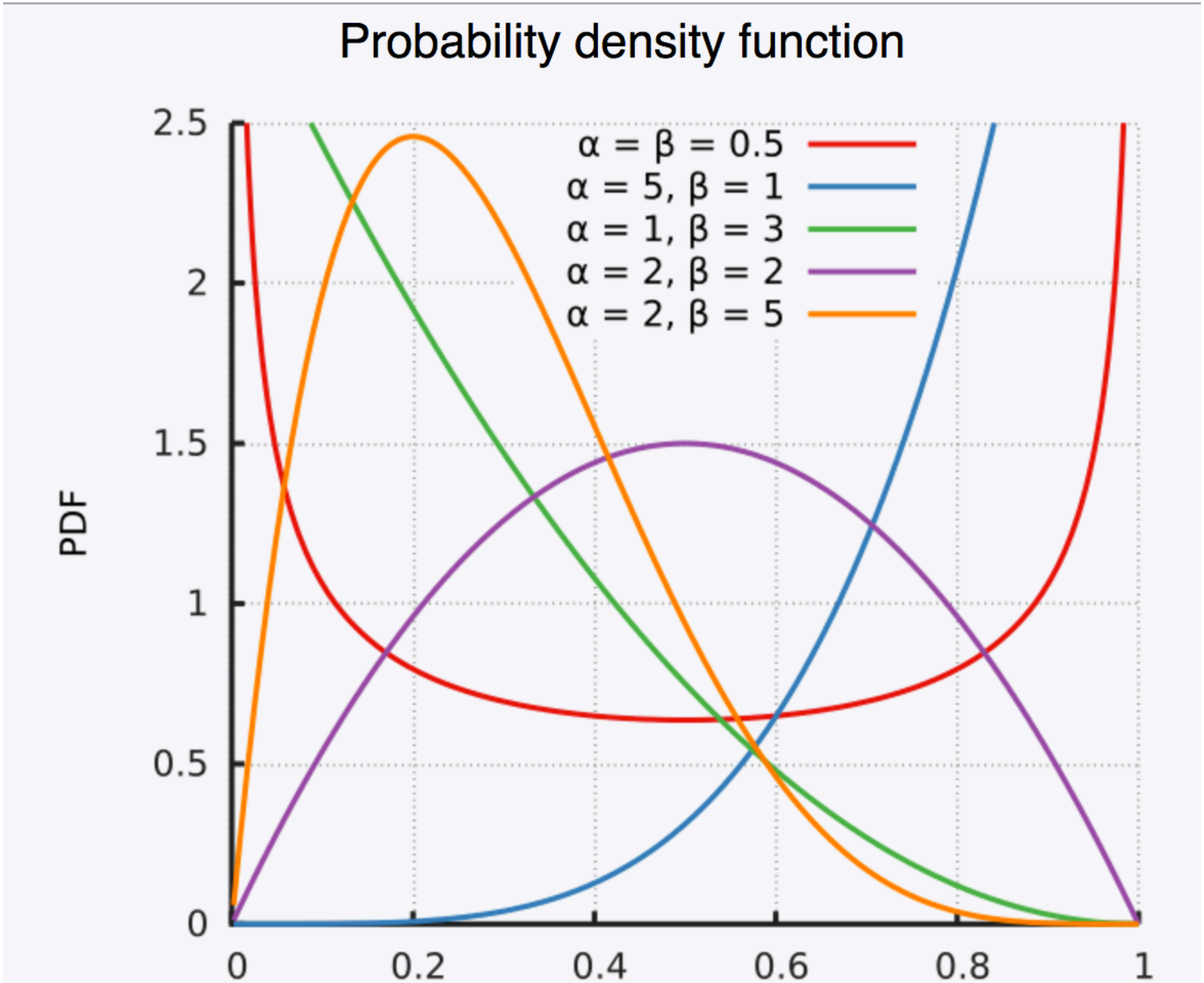
About your homework…

Beta Distribution

Study it in detail - Homework

PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Beta


Probability density function
PDF
α = β = 0.5
α = 5, β = 1
α = 1, β = 3
α = 2, β = 2
α = 2, β = 5
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

Review: Probability of an Event

- p(A) denotes the probability that the event A is true.
- For example:
- A = a logical expression "it will rain tomorrow"

We require that $0 \leq p(A) \leq 1$.

p(A) = 0 means the event definitely will not happen

p(A) = 1 means the event definitely will happen

$p(\overline{A})$ denotes the probability of the event not A

$$p(\overline{A}) = 1 - p(A)$$

We also write:

A=1 to mean the event A is true.

A=0 to mean the event A is false.

Review: Fundamental Rules

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned}$$

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3) \ldots p(X_D|X_{1:D-1})$$

- **Independence (or unconditionally independent** or **marginally independent)** denoted X ⊥ Y:

$$X \perp Y \iff p(X, Y) = p(X)p(Y)$$

- **Conditional Independence**

$$X \perp Y | Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Theorem: X ⊥ Y |Z *iff there exist function* g *and* h *such that*

$$p(x, y|z) = g(x, z)h(y, z)$$

$$\textit{for all } x, y, z \textit{ such that } p(z) > 0.$$

The **conditional probability** of event A, given that event B is true:

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

Bayes rule:

$$p(X=x|Y=y) = \frac{p(X=x,Y=y)}{p(Y=y)} = \frac{p(X=x)p(Y=y|X=x)}{\sum_{x'} p(X=x')p(Y=y|X=x')}$$

Example: medical diagnosis

- Suppose I did a medical test for breast cancer, called a **mammogram.** If the test is positive, what is the probability I have cancer? *(here y=1 means cancer is true, and x=1 means test is positive).*
- Suppose I have cancer, the test will be positive with probability 0.8. I.e. $p(x = 1|y = 1) = 0.8$.
- If I conclude therefore 80% likely I have cancer. ***True or False?***
- **False!**
- It ignores the prior probability of having breast cancer, which fortunately is quite low:
- $p(y = 1) = 0.004$

Using Byes Rule

$$p(y=1|x=1) = \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)}$$
$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Where 1) p(y = 0) = 1 − p(y = 1) = 0.996 .

2) Take into account the fact that the test may be a false positive or false alarm. With current screening technology:
p(x = 1 | y = 0) = 0.1

In other words, if I test positive, I only have about a 3% chance of actually having breast cancer!

Generative classifier

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{p(y=c|\boldsymbol{\theta})p(\mathbf{x}|y=c,\boldsymbol{\theta})}{\sum_{c'} p(y=c'|\boldsymbol{\theta})p(\mathbf{x}|y=c',\boldsymbol{\theta})}$$

This is called a **generative classifier**, since it specifies how to generate the data using the class- conditional density p(x|y = c) and the class prior p(y = c).

Change Gear to **The Generalized Linear Models (GLMs)**

Prof. Weiqing Gu
Harvey Mudd College

Summer 2017

https://math189su17.github.io/project.html

What is the Generalized Linear Models?

Linear Model

$Y = mX + b$ → $Y = \theta_0 + \theta_1 X_1$

X_i= house features
Y = predicted house price

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n$$

$$Y = \mathbf{X}^T \theta$$

Let $X_0 = 1$

(General) Linear Models

1. Extend predicted value to be vector valued. E.g. Y_1= price, Y_2 = how many people buy houses with the given the same features (X_1, X_2, ..., X_n)
-> **Multivariable regression**

2. Extend **X** to "catogrical". X_i = values of i^{th} category

3. Extend to Polynomial fitting:
$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + ... + \theta_n X^n$
It is still linear with respect to θ_i's.

Generalized Linear Models

Using hypothesis related to exponential family: the major part of it is an exponential of something, that something is a **Linear Model!**

(General) Linear Models

X_is are measured independent variables, may be continuous, may be categorical Or may be a mixture.

X	→	New X
1	→	1
2	→	0

X	→	New X_1		New X_2
1	→	1	→	0
2	→	0	→	1
3	→	0	→	0

Here we have 3 categories. The 3rd one with entries all 0, called the reference category.

Y is a measured dependent variable

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n + \epsilon$$

$$\mathbf{x}^T\theta$$

Residual/Erro term.

Regression weights, or parameters of the linear model, each assesses the feature/factor, X_i's contribution to predict the value of dependent variable Y. Note $x_0 = 1$. If all X_i=0, we will predict that the value Y to be θ_0 .

Story: How to predict Y from the knowledge of X_is?

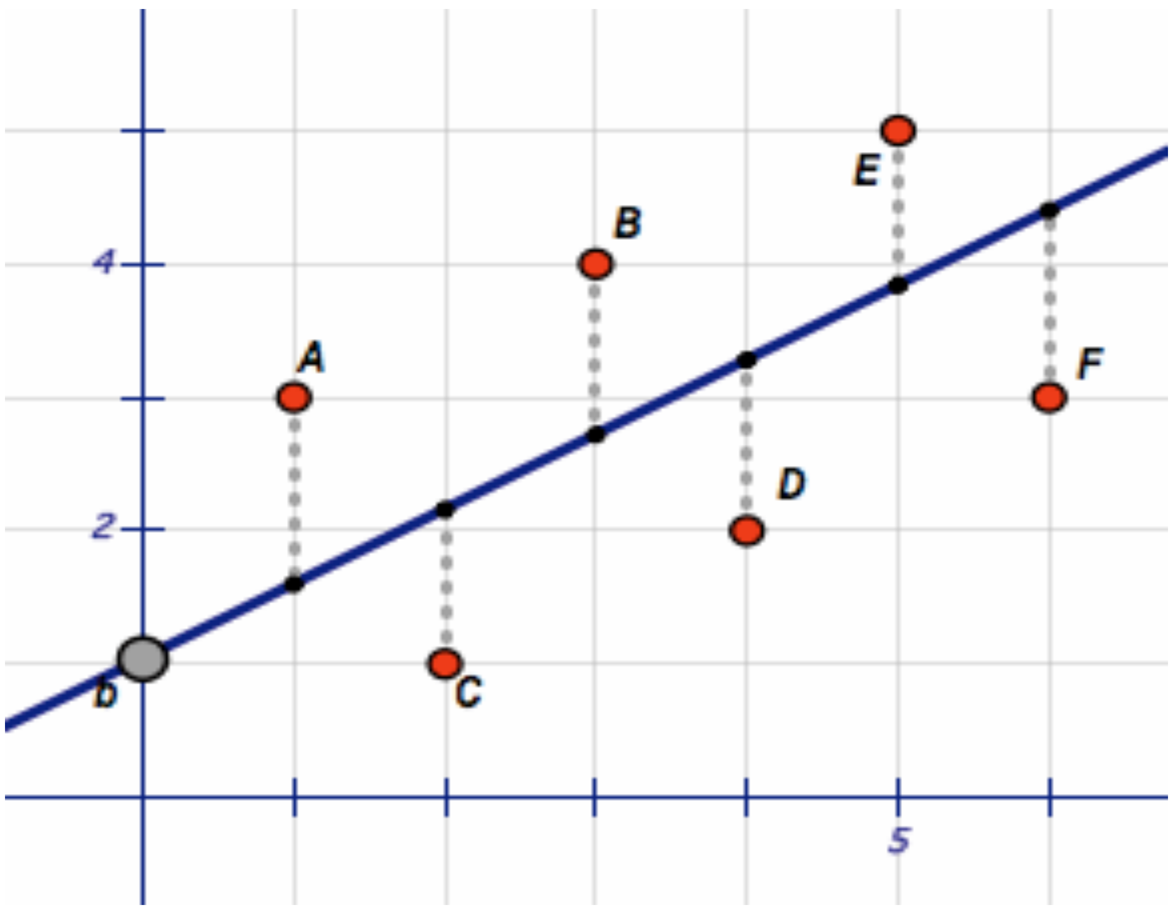
$\mathbf{x}^T\theta$ = the estimation of Y. It may not be accurate, too high, or too low.

ϵ = what can not be predicted from the knowledge of $\mathbf{x}^T\theta$. $\epsilon = Y - \mathbf{x}^T\theta$

The linear model answer the following questions:

- How do these independent factors (X_1, X_2, …, X_n) predict a single dependent variable (Y_i)?
- What is the best predictor of Y_i given measured X_is?
- Note for each Y_i, there is set of best weights.

$$(Y_1, Y_2) = (\mathbf{X}^T\theta_1, \mathbf{X}^T\theta_2) = \mathbf{X}^T(\theta_1, \theta_2)$$



Recall: For our linear model: $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)},$

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right).$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right).$$

$$y^{(i)} \mid x^{(i)}; \theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2).$$

Given X (the design matrix, which contains all the $x^{(i)}$'s) and θ, what is the distribution of the $y^{(i)}$'s? The probability of the data is given by $p(\vec{y}|X;\theta)$. This quantity is typically viewed a function of $\vec{y}$ (and perhaps X), for a fixed value of θ. When we wish to explicitly view this as a function of θ, we will instead call it the **likelihood** function:

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y}|X;\theta).$$

$$X = \begin{bmatrix} — (x^{(1)})^T — \\ — (x^{(2)})^T — \\ \vdots \\ — (x^{(m)})^T — \end{bmatrix}.$$

$$X\theta - \vec{y} = \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Recap how we find the maximum—This gives a general method called Maximum Likelihood Estimation.

- Obtain the likelihood

$$L(\mu) = f(y_1)\ldots f(y_n)$$

- Log it – to make it easier & fast in calculation. Keep the advantage of the linear predictor.

$$\ln L(\mu)$$

- Differential and set the derivative equal to 0.

$$\frac{d}{d\mu}\ln L(\mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \ldots$$

- Check it is a maximum: $\frac{d^2}{d\mu^2}\ln L(\mu) < 0 \quad \Rightarrow \quad \max$

Find parameters for the GLMs

- Obtain a likelihood function
- Log it to make it easier in differentiate
- Use the link function to replace the means resulting a function in the parameters.
- Differentiate with respect to the parameters and set the derivatives all to zero and solve for the optimal parameters.

Let's Derive

A GLM using

Multinomial distributions which we have shown that they exponential family distributions.

Recall: Generally an experiment with m outcomes with respective probabilities $p_1, p_2, \ldots, p_m$ is performed n times independently.

Let x_i = # of times outcome i appears, $i=1,2,\ldots,m$

Then $P(x_1=k_1, x_2=k_2, \ldots, x_m = k_m) = ?$

- Work out details with the students on the board.

Examples of

Generalized Linear Models (GLMs)

- Use GLMs and ***exponential family to get Softmax Regression***.
- *Recall: What is an exponential family?* A class of distributions is in the exponential family if

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta))$$

- η = the natural parameter (or the canonical parameter) of the distribution
- T(y) = the sufficient statistic (often T(y) = y)
- a(η) is the log partition function.

The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure the distribution p(y; η) sums/integrates over y to 1.

Let T, a and b fixed and let the parameter η vary, then it defines a family of distribution.
i.e. We get different distributions within this family.

We saw

Bernoulli distributions are exponential family distribution.

- Work out details with the students on the board.

Gaussian distributions are exponential family distribution.

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \cdot \exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

Compare:

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

We get:

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2). \end{aligned}$$

Example of Constructing GLMs

Note: you need to know which distribution models what kind of problems

(Reading assignment)

- Suppose you want to build a model to estimate the number (y) of customers arriving in your store in any given hour, based on certain features x such as store promotions, recent advertising, weather, day-of-week, etc.
- We know that the Poisson distribution usually gives a good model for numbers of visitors.
- Knowing this, how can we come up with a model for this problem?
- Fortunately, the Poisson is an exponential family distribution, so we can apply a Generalized Linear Model (GLM). *(Homework or exam problem?)*
- Lots of known distributions are exponential families.
- Here, we will describe a method for constructing GLM models for problems such as these.

Assumptions for Generalized Linear Models

- In generally, consider a classification or regression problem where we would like to predict the value of some random variable y as a function of x.
- To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given x and about our model:
- **1. $y \mid x; \theta \sim$ Exponential Family(η)**. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.
- **2**. Given x, our goal is to predict the expected value of T(y) given x. Since often T(y) = y, so this means we would like the prediction **h(x) output by our learned hypothesis h to satisfy h(x) = E[y|x].** (Note that this assumption is satisfied in the choices for $h_\theta(x)$ for both logistic regression and linear regression. For instance, in logistic regression, we had

$h_\theta(x) = p(y = 1|x; \theta) = 0 \cdot p(y = 0|x; \theta) + 1 \cdot p(y = 1|x; \theta) = E[y|x; \theta]$.)

- **3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$**. (Or, if η is vector-valued, then $\eta_i = \theta_i^T x$.)

Examples: Least square and Logistic regression are GLM family of models

$$
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \mu \\
&= \eta \\
&= \theta^T x.
\end{aligned}
$$

$$
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \phi \\
&= 1/(1+e^{-\eta}) \\
&= 1/(1+e^{-\theta^T x})
\end{aligned}
$$

Given that y is binary-valued, it therefore seems natural to choose the Bernoulli family of distributions to model the conditional distribution of y given x. In our formulation of the Bernoulli distribution as an exponential family distribution, we had $\phi = 1/(1 + e^{-\eta})$. Furthermore, note that if $y|x; \theta \sim \text{Bernoulli}(\phi)$, then $E[y|x; \theta] = \phi$.

Softmax Regression

- Let's look at another example of a GLM. Consider a classification problem in which the response variable y $\in \{1, 2, . . . , k\}$.
- For example, rather than classifying email into the two classes spam or not-spam—which would have been a binary classification problem— this time we want to classify it into four classes, such as spam, family-mail, friends-mail, and work-related mail. The response variable is still discrete, but can now take on more than two values. We will thus model it as distributed according to a multinomial distribution.

Details of Softmax Regression

- Work out details with the students on the board.

Today we also learn:

Schur Complement

- This is related how we triage data and solve a smaller problem involving big data first.
  - Smaller system to solve
  - Smaller matrix to invert
  - The process can be iterated to make the problem to a smaller and smaller size. (This is very powerful for dealing with big data. This is one of the dimension reduction methods.)
- It is also very important for study the Conditional Gaussian distribution.
- Work out details with the students on the board.

What is a conditional distribution?

- A conditional distribution is a probability distribution for a sub-population.
- In other words, it shows the probability that a randomly selected item in a sub-population has a characteristic you're interested in.
- For example, if you are studying eye colors (the population) you might want to know how many people have blue eyes (the sub-population).

Conditional Distribution
Discrete example

		Eye Color			
		Blue	Brown	Green/Other	Total
Gender	Male	15	20	8	43
	Female	5	25	7	37
	Total	20	45	15	80

e.g. We restrict to only on Blue eyes, the conditional distribution is Male:15 and Femaie:5 . This is called a conditional distribution.

Conditional Distribution (continuous)

If N-dimensional $\mathbf{x}$ is partitioned as follows

Later!

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

and accordingly $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned as follows

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

then the distribution of $\mathbf{x}_1$ conditional on $\mathbf{x}_2 = \mathbf{a}$ is multivariate normal $(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{a} - \boldsymbol{\mu}_2)$$

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$$ ← the Schur complement of $\boldsymbol{\Sigma}_{22}$ in $\boldsymbol{\Sigma}$

Back up slides

Note: Polynomial data fitting is also a linear model, also will be resulted in the normal equation

x_i	y_i
1	1
2	5
3	8
4	17
5	16

We always get the same normal equation!

$$\begin{bmatrix} 1^2 & 1 & 1 \\ 2^2 & 2 & 1 \\ 3^2 & 3 & 1 \\ 4^2 & 4 & 1 \\ 5^2 & 5 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 8 \\ 17 \\ 26 \end{bmatrix}.$$

So, a good fit to the data is to find a, b, and c such that $y(x) = ax^2 + bx + c$ is "closest" to the data. In the least squares sense the means for $r_i = y_i - y(x_i) = y_i - (a\,x_i^2 + b\,x_i + c)$.

Same geometric argument works to get the normal equation!

When we have polynomials with multi-variables, the size of the X^TX can be very large.