MACHINE LEARNING

Foundations: Probability

Last Update: 17th October 2022

Prof. Dr. Shadi Albarqouni

Director of Computational Imaging Research Lab. (Albarqouni Lab.)
University Hopsital Bonn | University of Bonn | Helmholtz Munich

STRUCTURE

1. Intro. to Probability
2. Univariate Models
3. Multivariate Models

INTRO. TO PROBABILITY

Warm-up Example: Fish bowls

posterior | liklihood | prior

Given two bowls, where

- in bowl-1 there are 30 red fishes and 10 blue fishes while
- in bowl-2 there 20 red fishes and 20 blue fishes,

and you catched a red fish without looking, what is the probability that the fish came from bowl-1?

CREATED BY VECTORPORTAL.COM

Inspired by the example in this blog https://medium.com/@canerkilinc/bayess-theorem-a0f6e3537278

WHAT IS PROBABILITY?

Probability theory is nothing but common sense reduced to calculation. – (Pierre Laplace, 1749-1827)

- Frequentist interpretation: probabilities represent long run frequencies of events.
- Bayesian interpretation: probability is used to quantify our uncertainty or ignorance about something; that can happen multiple times.
- Real-life applications include but not limited to; Weather Forecasting, Politics, Insurance among others.

PROBABILITY AS AN EXTENSION OF LOGIC

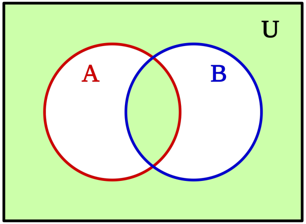
The expression $Pr(A)$ denotes the probability with which you believe event A is true. We require that $0 < Pr(A) < 1$, where $Pr(A) = 0$ means the event definitely will not happen, and $Pr(A) = 1$ means the event definitely will happen.

Joint probability: $Pr(A \wedge B) = Pr(A, B)$

Union probability:
$Pr(A \vee B) = Pr(A) + Pr(B) - Pr(A \wedge B)$

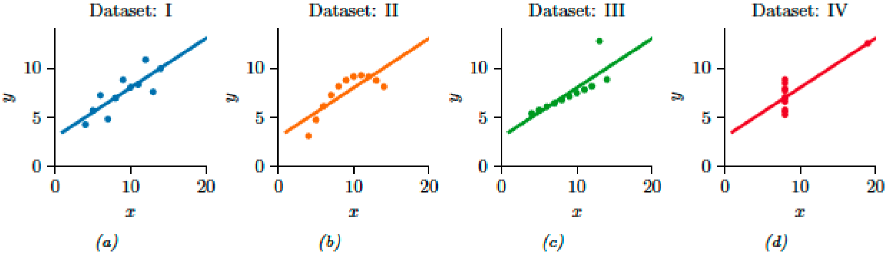
Conditional probability: $Pr(B|A) \triangleq \frac{Pr(A,B)}{Pr(A)}$

Indpendence of events:
$Pr(A, B) = Pr(A)Pr(B)$ iff $A \perp B$



LIMITATIONS OF SUMMARY STATISTICS

Anscombe's quartet (Code)

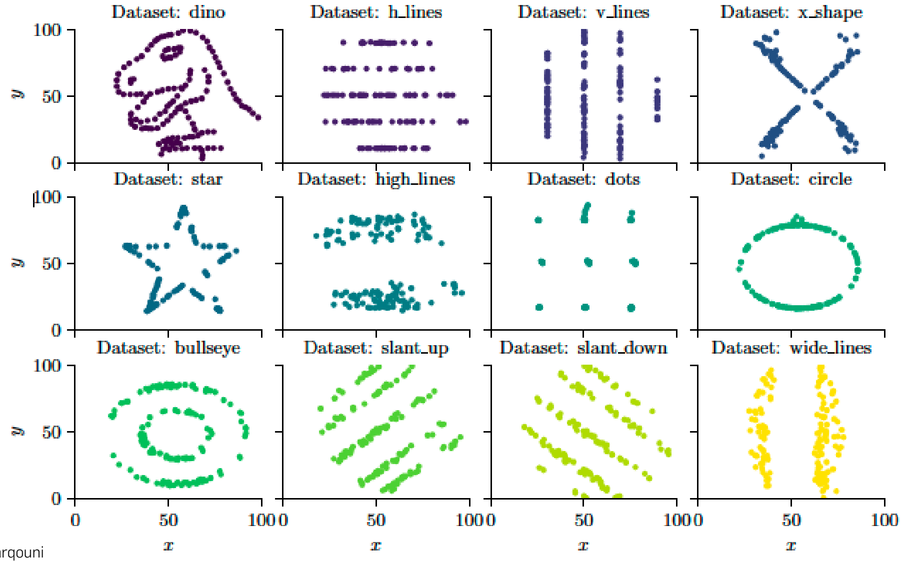


Compute the expected value $\mathbb{E}[\cdot]$ and variance $\mathbb{V}[\cdot]$ of the random variables x and y

Compute the correlation coefficient ρ

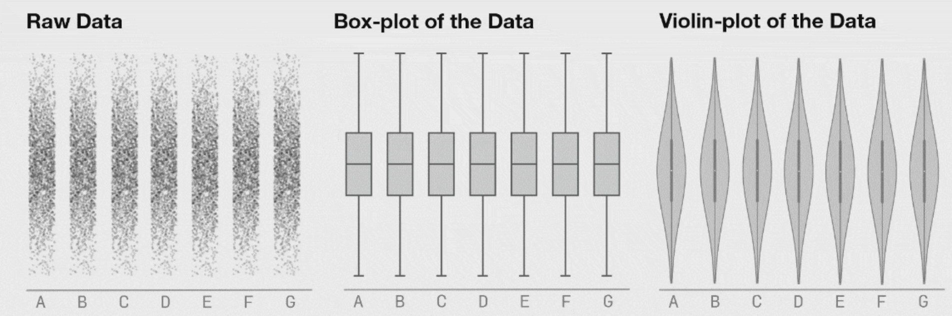
Report your observation

Datasaurus Dozen (Code)



VISUALIZATION VS. STATISTICS

Box plot vs. violin plot in Python (Code)



limitations of visualization?
features beyond statistics!

Source:https://www.autodesk.com/research/publications/same-stats-different-graphs

BAYES' THEOREM

Bayes's theorem is to the theory of probability what Pythagoras's theorem is to geometry. — Sir Harold Jeffreys, 1973

Bayes' Theorem

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

- The term $p(H)$ represents what we know about possible values of hypotheses H before we see any data/observations; this is called the prior distribution.
- The term $p(Y = y|H = h)$ represent the probability at a point corresponding to the actual observations, y which is called the likelihood.
- The term $p(Y = y)$ is known as the marginal likelihood and computed as $\sum_{h' \in \mathcal{H}} p(H = h')p(Y = y|H = h')$
- The term $p(H = h|Y = y)$ represent the posterior distribution

Example: Fish bowls -- two more examples in Sec. 2.3.1

posterior | liklihood | prior

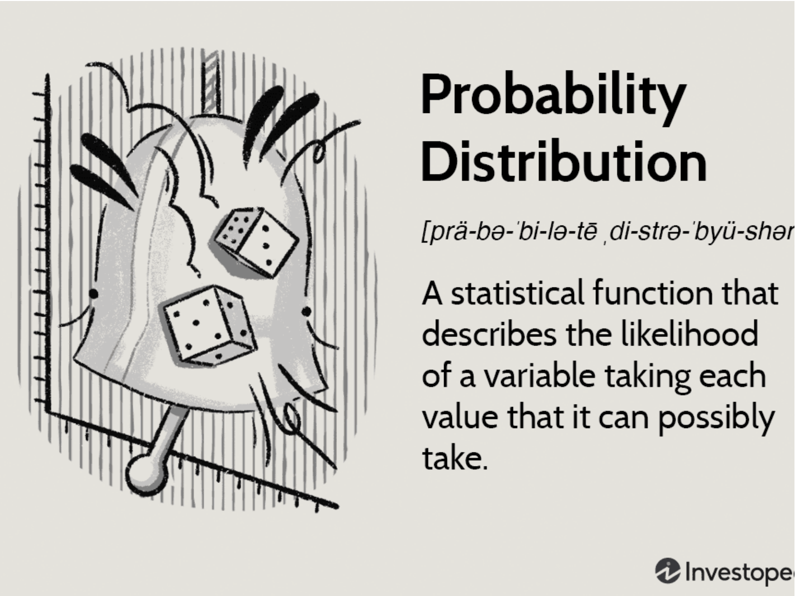
Given two bowls, where

- in bowl-1 there are 30 red fishes and 10 blue fishes while
- in bowl-2 there 20 red fishes and 20 blue fishes,

and you catched a red fish without looking, what is the probability that the fish came from bowl-1?

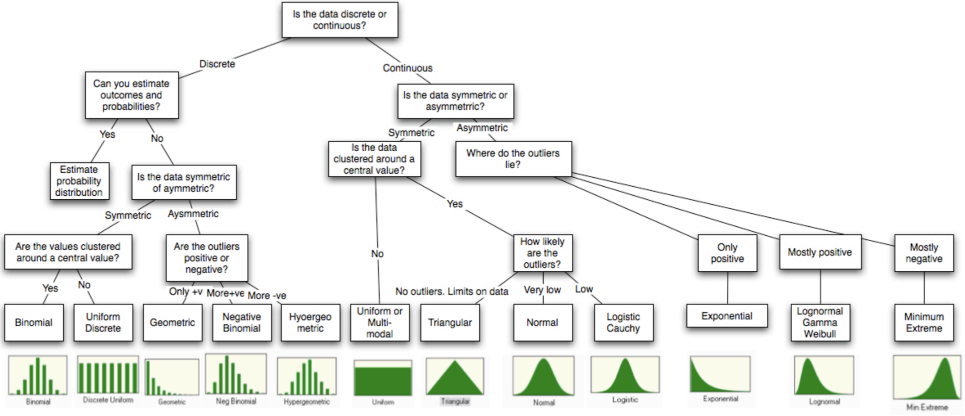
Inspired by the example in this blog https://medium.com/@canerkilinc/bayess-theorem-a0f6e3537278

UNIVARIATE MODELS



Source:**https://www.investopedia.com/terms/p/probabilitydistribution.asp**

DIFFERENT TYPES OF DISTRIBUTIONS



Source: Fig. 6A.15 from https://pages.stern.nyu.edu/ adamodar/pdfiles/papers/probabilistic.pdf

BERNOULLI AND BINOMIAL DISTRIBUTIONS

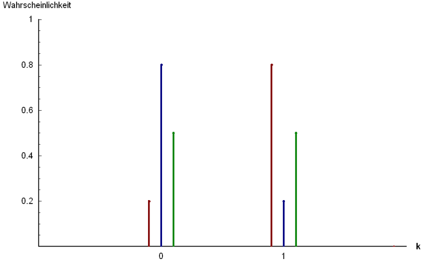
Bernoulli distribution (pmf)

$$\text{Ber}(y|\theta) = \begin{cases} 1-\theta & \text{if } y = 0 \\ \theta & \text{if } y = 1 \end{cases}$$

It can be written as $\text{Ber}(y|\theta) \triangleq \theta^y(1-\theta)^{1-y}$ where θ is the probability of event $y = 1$.

$$\mathbb{E}[y] = \sum_{y=0}^{1} y\text{Ber}(y|\theta) = \theta$$

$$\mathbb{V}[y] = \sum_{y=0}^{1}(y - \mathbb{E}[y])^2\text{Ber}(y|\theta) = \theta(1-\theta)$$



Source:
https://en.wikipedia.org/wiki/Bernoulli_distribution

Bionomial distribution (pmf)

$$\mathrm{Bin}(s|N,\theta) = \binom{N}{s}\theta^s(1-\theta)^{N-s}$$

where $\binom{N}{k} = \frac{N!}{(N-k)!k!}$,

θ is the probability of event $y = 1$,

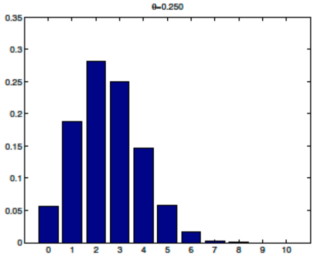
N is the number of trials, and

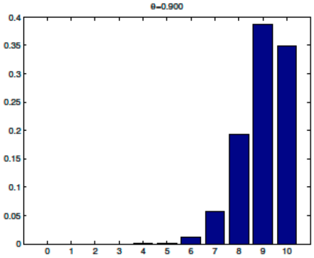
$s \triangleq \sum_{n=1}^{N} \mathbb{I}(y_n = 1)$ is the total number of an event $y = 1$.

Compute $\mathbb{E}[y]$ and $\mathbb{V}[y]$

Special case:

$\mathrm{Bin}(s|N,\theta) = \mathrm{Ber}(y|\theta) \triangleq \theta^y(1-\theta)^{1-y}$ when $N = 1$.





Play with the Code – take Castania as an example

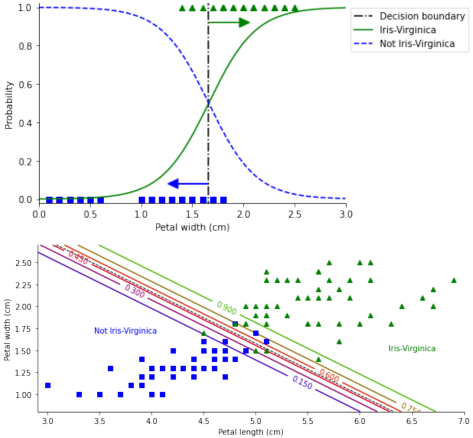
Example: classifying Iris flowers (Code)

Sigmoid (logistic) function | heaviside step function | Self-reading Sec. 2.5

Given some inputs $x \in \mathcal{X}$ and a mapping function $f(\cdot)$ that predict a binary variable $y \in \{0, 1\}$, write the conditional probability distribution $p(y|x, \theta)$:

$$p(y|x, \theta) = \text{Ber}(y|f(x; \theta))$$

To avoid the requirement that $0 < f(x; \theta) < 1$, we use the following model $p(y|x, \theta) = \text{Ber}(y|\sigma(f(x; \theta)))$, where $\sigma(a) = \frac{1}{1+\exp^{-a}}$ is the sigmoid function with $a = f(x; \theta)$.



UNIVARIATE GAUSSIAN (NORMAL) DISTRIBUTION

The most widely used distribution of real-valued random variables $y \in \mathbb{R}$ is the Gaussian distribution, also called the normal distribution.

Gaussian distribution (pdf)

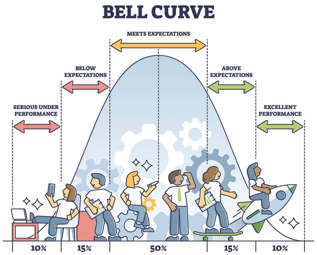
$$p(y) = \mathcal{N}(y|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

where $\sqrt{2\pi\sigma^2}$ is the normalization constant.

$$\mathbb{E}[y] = \int_{\mathcal{Y}} y p(y) = \mu$$

$$\mathbb{V}[y] = \int_{\mathcal{Y}} (y - \mathbb{E}[y])^2 p(y) = \sigma^2$$

Special case: $\mathcal{N}(y|0, 1)$ is the standard normal distribution



Source: **https://www.simplypsychology.org/normal-distribution.html**

Why is it so widely used?

- two parameters easy to interpret
- central limit theorem; sum of i.i.d random variables –> gaussian distribution
- makes the least number of assumptions (max. entropy) –> good default choice
- simple mathematical form to implement

Example: regression (Code)

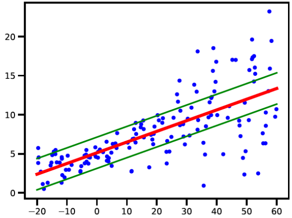
linear regression | homoscedastic regression | heteroscedastic regression | Softplus

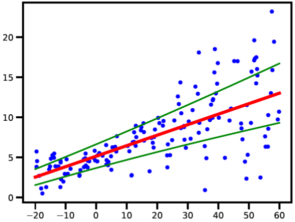
Given some inputs $x \in \mathcal{X}$ and a mapping function $f(\cdot)$ that predict the response $y \in \mathcal{Y}$, write the conditional probability distribution $p(y|x,\theta)$ as a conditional guassian distribution.

$$p(y|x,\theta) = \mathcal{N}(y|f_\mu(x;\theta), f_\sigma(x;\theta)^2)$$

where $f_\mu(x;\theta) \in \mathbb{R}$ predicts the mean, and $f_\sigma(x;\theta) \in \mathbb{R}_+$ predicts the variance.

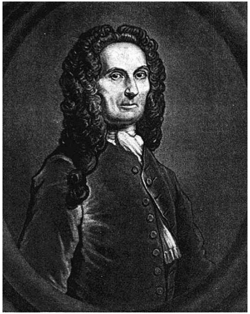
- homoscedastic regression: The variance is independent of the input; $\mathcal{N}(y|\mathbf{w}^T\mathbf{x} + b, \sigma^2)$
- heteroscedastic regression The variance is a function of the input; $\mathcal{N}(y|\mathbf{w}_\mu^T\mathbf{x} + b, \sigma_+(\mathbf{w}_\sigma^T\mathbf{x}))$





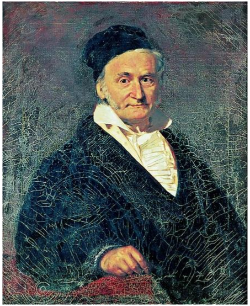
FUN FACTS

"The fundamental nature of the Gaussian distribution and its main properties were noted by Laplace when Gauss was six years old; and the distribution itself had been found by de Moivre before Laplace was born" – Jaynes



Abraham de Moivre (1667 - 1754)

Pierre Simon Laplace (1749 - 1827)



Carl Friedrich Gauss (1777 - 1855)

DIRAC DELTA FUNCTION

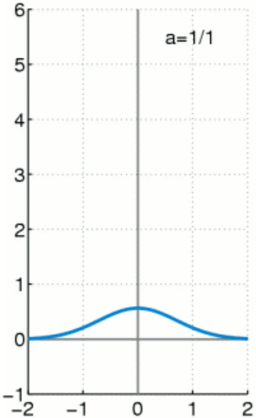
As the variance σ^2 in the Gaussian distribution goes to zero, the distribution approaches an infinitely narrow, but infinitely tall, "spike" at the mean $p(y) \triangleq \lim_{\sigma \to 0} \mathcal{N}(y|\mu, \sigma^2) \to \delta(y - \mu)$

Dirac delta function

$$\delta(x) = \begin{cases} +\infty & \text{if} x = 0 \\ 0 & \text{otherwise} \end{cases},$$

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

Sifting property: $\int_{\infty}^{-\infty} f(y)\delta(x-t)dy = f(x)$

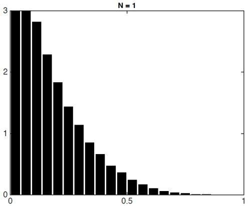


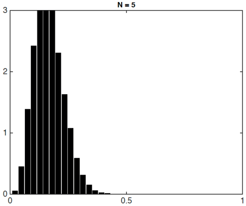
Source: https://https://commons.wikimedia.org/

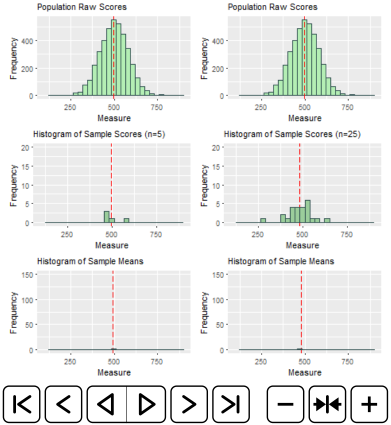
CENTRAL LIMIT THEOREM (CODE)

Definition

The distribution of the sum of N independent and identically distributed (i.i.d) random variables $X_n \sim p(X)$, e.g., $S_{N_\mathcal{D}} = \sum_{n=1}^{N_\mathcal{D}} X_n$, converges to a standard normal distribution where $\bar{X} = S_N/N$ is the sample mean.





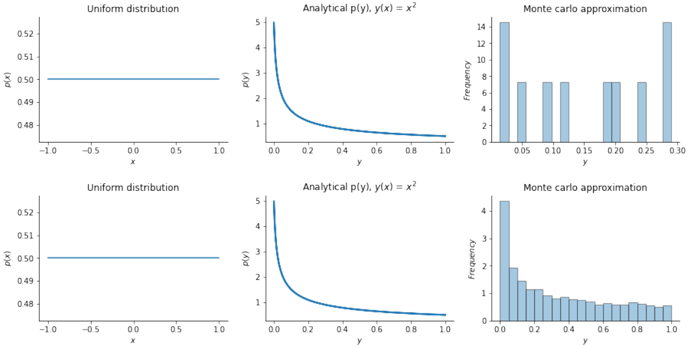


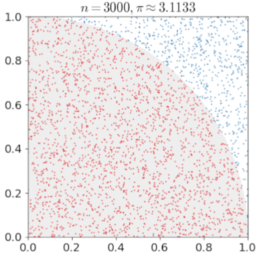
Source: developed by William Arloff
https://you.stonybrook.edu/banderson/statistics/

MONTE CARLO APPROXIMATION (CODE)

Definition

It is a common approach to recover the underlying distribution $p(y)$ where $y = f(x)$ by drawing many samples from a random number generator $p(x)$

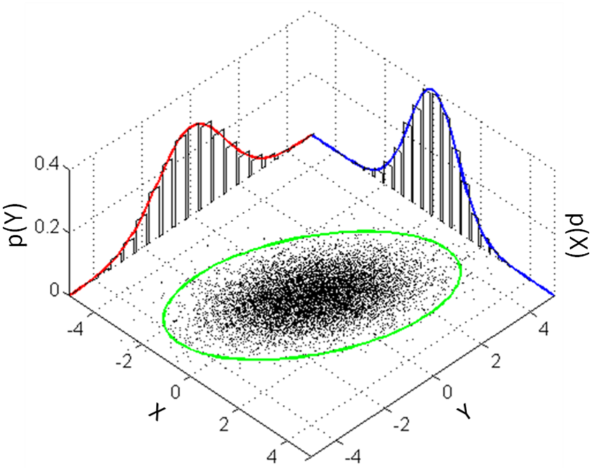




Source:**https://en.wikipedia.org/wiki/Monte_ Carlo_method**

MULTIVARIATE MODELS

UNIVARIATE VS. MULTIVARIATE RANDOM VARIABLES



MULTIVARIATE GAUSSIAN (NORMAL) DISTRIBUTION

The most widely used joint probability distribution for continuous random variables is the multivariate Gaussian or multivariate normal (MVN).

Multivariate Gaussian distribution (pdf)

$$\mathcal{N}(\mathbf{y}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where

$\mathbb{E}[\mathbf{y}] = \mu$ is the mean vector,
$\mathrm{Cov}[\mathbf{y}] \triangleq \mathbb{E}\left[(\mathbf{y}-\mathbb{E}[\mathbf{y}])(\mathbf{y}-\mathbb{E}[\mathbf{y}])^T\right]$ is the covaraince matrix, and
$Z = (2\pi)^{D/2}|\Sigma|^{1/2}$ is the normalization constant
$\mathbb{E}[\mathbf{y}\mathbf{y}^T] = \Sigma + \mu\mu^T$

Example: bivariate Gaussian distribution (Code)

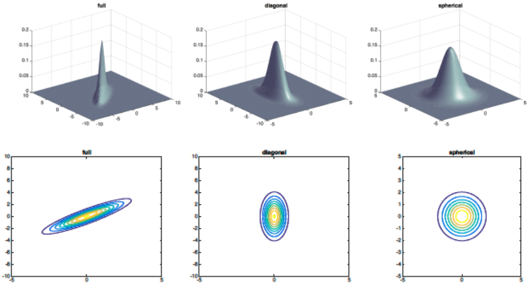
$$\mathcal{N}(\mathbf{y}|\mu,\Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)$$

where

$$\mu = (\mu_1, \mu_2)^T,$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$ with

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\mathbb{V}[Y_1]\mathbb{V}[Y_2]}} = \frac{\sigma_{12}^2}{\sigma_2\sigma_2}$$ as a correlation coefficient



- What is the difference between full, diagonal, and spherical covariance matrices?

MARGINALS AND CONDITIONALS OF AN MVN

Suppose $\mathbf{y} = (\mathbf{y_1}; \mathbf{y_2})$ is jointly Gaussian with parameters $\mu = (\mu_1, \mu_2)^T$, and $\Sigma = \begin{pmatrix} \Sigma_{11}^2 & \Sigma_{12}^2 \\ \Sigma_{21}^2 & \Sigma_{22}^2 \end{pmatrix}$.

The marginals are given by:

$$p(\mathbf{y_1}) = \mathcal{N}(\mathbf{y_1} | \boldsymbol{\mu_1}, \Sigma_{11})$$

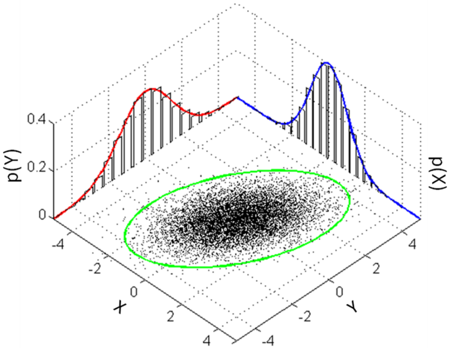
$$p(\mathbf{y_2}) = \mathcal{N}(\mathbf{y_2} | \boldsymbol{\mu_2}, \Sigma_{22})$$

The posterior conditional is given by:

$$p(\mathbf{y_1} | \mathbf{y_2}) = \mathcal{N}(\mathbf{y_1} | \mu_{1|2}, \Sigma_{1|2}) \text{ where}$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y_2} - \boldsymbol{\mu_2})$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$



Conditioning on a 2d Gaussian

missing value imputation | multiple imputation

Given a set of 2d points centered around zero mean with a unit standard deviation for σ_1 and σ_2 and a correlation coffiecient of 0.7, what would be the expected value of y_1 given $y_2 = 1$? What happens if $\rho = 0$? Could you tell whether the covariance matris is full, diagonal, or spherical?

The following formuls might be helpful to solve the problem:

Mean: $\mu = (\mu_1, \mu_2)$

Covariance matrix: $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

Marginal distribution: $p(y_1) = \mathcal{N}(y_1|\mu_1, \sigma_1^2)$

Conditional distribution: $p(y_1|y_2) = \mathcal{N}\left(y_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(y_2 - \mu_2), \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$

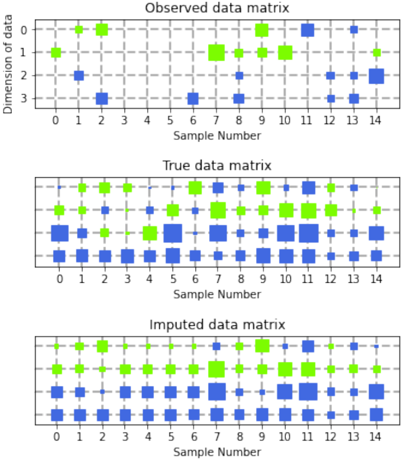
The answer is . . .

Example: Imputing missing values (Code)

missing value imputation | multiple imputation | Hinton diagram

Given 15 vectors sampled from a 4 dimensional Gaussian, infer the missing values h given the observed ones v.

- compute the mean μ and covariance matrix Σ given the observed data
- compute the marginal distribution of each missing value $p(y_{n,h}|y_{n,v},(\mu,\Sigma))$
- compute the posterior mean $\bar{y_{n,i}} = \mathbb{E}[y_{n,i}|\mathbf{y_{n,v}},(\mu,\Sigma)]$



Questions