Probability Distributions

Sadegh Eskandari
Department of Computer Science, University of Guilan

Probability Distributions: Introduction

- **Remember:** Probability theory provides a consistent framework for the quantification and manipulation of uncertainty in data

- **Density Estimation**: given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations for a random variable $\mathbf{x}$, the goal is to model the probability distribution $p(\mathbf{x})$.

- We will assume that the data points are independent and identically distributed (iid).

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x_n})$$

Density Estimation

- **Parametric**
  - Selecting a common distribution and estimating the parameters for the density function from the data
  - binomial and multinomial distributions for discrete random variables
  - Gaussian distribution for continuous random variables.
  - Parameter estimation procedure: maximum likelihood, Bayesian method
- **Non-Parametric**
  - Histograms, Nearest-Neighbours, Kernels

Binary Variables

Bernoulli Distribution

- Consider a single binary random variable $x \in \{0,1\}$
- For example, x might describe the outcome of flipping a coin, with $x = 1$ representing 'heads', and $x = 0$ representing 'tails'.
- The probability of $x = 1$ will be denoted by the parameter $0 \leq \mu \leq 1$ so that:

$$p(x = 1|\mu) = \mu$$

- And hence:

$$p(x = 0|\mu) = 1 - \mu$$

- Therefore, the probability distribution over x can be written in the form:

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

Binary Variables

Bernoulli Distribution

$$\mathbb{E}[x] = \sum_x x p(x)$$

$$= \sum_x x\mathrm{Bern}(x|\mu) = 0 \times \mathrm{Bern}(x = 0|\mu) + 1 \times \mathrm{Bern}(x = 1|\mu) = \mu$$

$$var(x) = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[(x - \mu)^2] = \sum_x (x - \mu)^2 p(x)$$

$$= (0 - \mu)^2 \mathrm{Bern}(x = 0|\mu) + (1 - \mu)^2 \mathrm{Bern}(x = 1|\mu)$$

$$= \mu^2(1 - \mu) + (1 - \mu)^2 \mu = \mu(1 - \mu)$$

Binary Variables

Bernoulli Distribution

- Now suppose we have a data set $\mathcal{D} = \{x_1, \dots, x_N\}$ of observed values of x
- We know that $\mathcal{D}$ is derived from a Bernoulli distribution
- But, we do not know the parameter μ
- So, we want to estimate μ using $\mathcal{D}$
- The maximum likelihood approach:

$$\mu_{ML} = \arg\max_{\mu} p(\mathcal{D}|\mu) = \arg\max_{\mu} \prod_{n=1}^{N} p(x_n|\mu) = \arg\max_{\mu} \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$= \arg\max_{\mu} \ln\left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right)$$

$$= \arg\max_{\mu} \left(\underbrace{\sum_{n=1}^{N}\{x_n \ln\mu + (1-x_n)\ln(1-\mu)\}}_{f(\mu)}\right)$$

$$\frac{\partial f}{\partial \mu} = 0 \quad \Longrightarrow \quad \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

m: the number of observations of $x = 1$

Binary Variables

Bernoulli Distribution

- Now suppose we flip a coin, say, 3 times and happen to observe 3 heads.
- Then $N = m = 3$ and $\mu_{ML} = 1$
- Then the maximum likelihood result would predict that all future observations should give heads!
- In fact this is an extreme example of the over-fitting associated with maximum likelihood.

- **Solution**: Bayesian Approach

Likelihood

Posterior Probability

Prior Probability

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu) \times p(\mu)}{p(D)}$$

Binary Variables

Bernoulli Distribution

Likelihood

Posterior Probability

Prior Probability

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu) \times p(\mu)}{p(D)}$$

- Step1:Likelihood function

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} = \mu^{\sum_{n=1}^{N} x_n}(1-\mu)^{N-\sum_{n=1}^{N} x_n} = \mu^m(1-\mu)^{N-m}$$

- Step2: Prior Probability
  - In this step we need to introduce a prior distribution $p(\mu)$ over the parameter μ. But how?
  - Conjugacy: As the likelihood function takes the form of powers of μ and $1-\mu$, then, if we choose a prior to be proportional to powers of μ and $1-\mu$, then the posterior distribution will have the same functional form as the prior.
  - Here we choose a prior called Beta distribution:

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b} \qquad var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du$$

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1) = 1$$

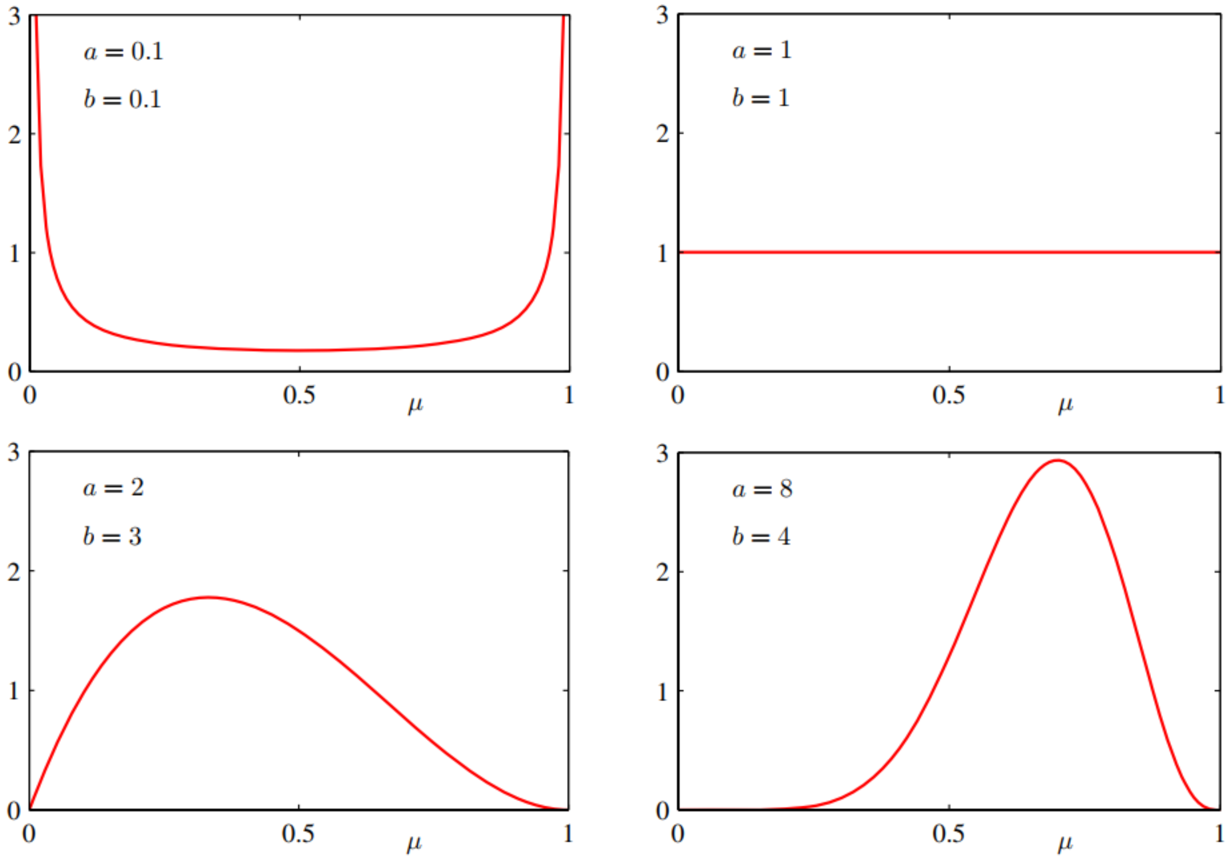
Proof: Homework

$$\Gamma(x) = x!, \forall x \in \mathbb{Z}$$

Binary Variables

Bernoulli Distribution

- Plots of $Beta(\mu|a,b)$ given by for various values of the hyperparameters a and b.



- Step3: Posterior Probability

$$p(\mu|m,l,a,b) = \frac{\mu^m(1-\mu)^l \times Beta(\mu|a,b)}{\int_0^1 \mu^m(1-\mu)^l \times Beta(\mu|a,b)d\mu}$$

$$= \frac{\Gamma(a+m+b+l)}{\Gamma(a+m)\Gamma(\mathrm{b}+\mathrm{l})}\mu^{a+m-1}(1-\mu)^{b+l-1}$$

- Sequential Learning: The posterior distribution can act as the prior if we subsequently observe additional data (applicable for big data).

Binary Variables

Bernoulli Distribution

- **Question**: How to predict the outcome of the next trial of x, given the observed data set $\mathcal{D}$?

$$p(x=1|\mathcal{D})$$

$$=\int_0^1 p(x=1,\mu|\mathcal{D})d\mu$$

$$=\int_0^1 p(x=1\ |\mu,\mathcal{D})\ \ p(\mu|\mathcal{D})\ d\mu$$

$$=\int_0^1 p(x=1\ |\mu)\ \ p(\mu|\mathcal{D})\ d\mu$$

$$=\int_0^1 \mu p(\mu|\mathcal{D})\ d\mu$$

$$=\mathbb{E}[\mu|\mathcal{D}]=\frac{m+a}{m+a+l+b}$$

Binary Variables

Binomial Distribution

- The number m of observations of $x = 1$, given that the data set has size N

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

- Where

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!\,m!}$$

$$\mathbb{E}[m] = \sum_{m=0}^{N} m\,Bin(m|N,\mu) = N\mu$$

$$var[m] = N\mu(1-\mu)$$

Proof: Homework

Binary Variables

Generalized Bernoulli Distribution

- Often, we encounter discrete variables that can take on one of K possible mutually exclusive states.

$$x \in \{s_1, s_2, \dots, s_K\} \quad \xrightarrow{\text{1-of-K encoding}} \quad \mathbf{x} \in \begin{Bmatrix} (1,0,0,\dots,0)^T \\ (0,1,0,\dots,0)^T \\ \vdots \\ (0,0,0,\dots,1)^T \end{Bmatrix}$$

- The random vector $\mathbf{x}$ can be described by K binary variables $x_1, x_2, \dots, x_K$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} \quad \text{Such that} \quad \sum_{k=1}^{K} x_k = 1$$

$$p(x_k = 1|\mu_k) = \mu_k$$

$$\mu_k \geq 0, \qquad \sum_{k=1}^{K} \mu_k = 1$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix}$$

Binary Variables

Generalized Bernoulli Distribution

- Now consider a data set $\mathcal{D}$ of N independent observations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$.
- So, we want to estimate the vector $\boldsymbol{\mu}$ using $\mathcal{D}$
- The Maximum Likelihood Approach

$$m_k = \sum_{n=1}^{N} x_{nk}$$

$$\boldsymbol{\mu}_{ML} = \arg\max_{\boldsymbol{\mu}} p(\mathcal{D}|\boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}} \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}) = \arg\max_{\mu_1,\dots,\mu_K} \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} = \arg\max_{\mu_1,\dots,\mu_K} \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} = \arg\max_{\mu_1,\dots,\mu_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$= \arg\max_{\mu_1,\dots,\mu_K} \ln\left(\prod_{k=1}^{K} \mu_k^{m_k}\right) = \arg\max_{\mu_1,\dots,\mu_K} \sum_{k=1}^{K} m_k \ln \mu_k$$

$$\text{s.t} \quad \sum_{k=1}^{K} \mu_k = 1$$

Lagrange

$$\mu_k^{ML} = \frac{m_k}{N}$$

Binary Variables

Generalized Bernoulli Distribution

Remember

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu) \times p(\mu)}{p(D)}$$

- The Bayesian Approach

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{m_k}$$

Conjugacy
(Dirichlet Distribution)

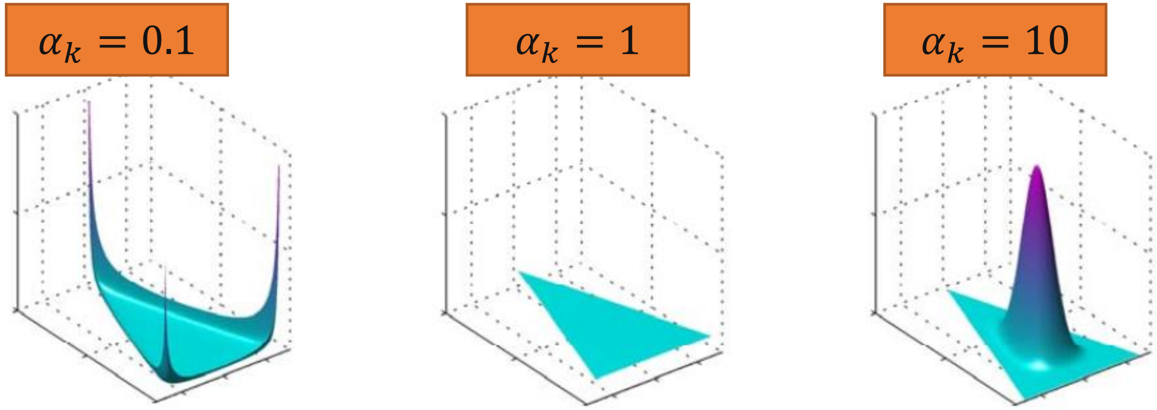
$$p(\boldsymbol{\mu}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$p(\boldsymbol{\mu}|\mathcal{D}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m})$$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\ldots\Gamma(\alpha_K + m_N)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

$$\mathbf{m} = (m_1, \ldots, m_K)^T$$



Dirichlet Distribution for $K = 3$

Binary Variables

Multinomial Distribution

- The joint distribution of the quantities $m_1, \dots, m_K$, conditioned on the parameters $\boldsymbol{\mu}$ and on the total number N of observations:

$$\mathrm{Mult}(m_1, m_2 \dots, m_K|\boldsymbol{\mu}) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

- Where

$$\binom{N}{m_1 m_2 \dots m_K} \equiv \frac{N!}{m_1!\, m_2! \dots m_K!}$$

$$\sum_{k=1}^{K} m_k = N$$

Gaussian Distribution

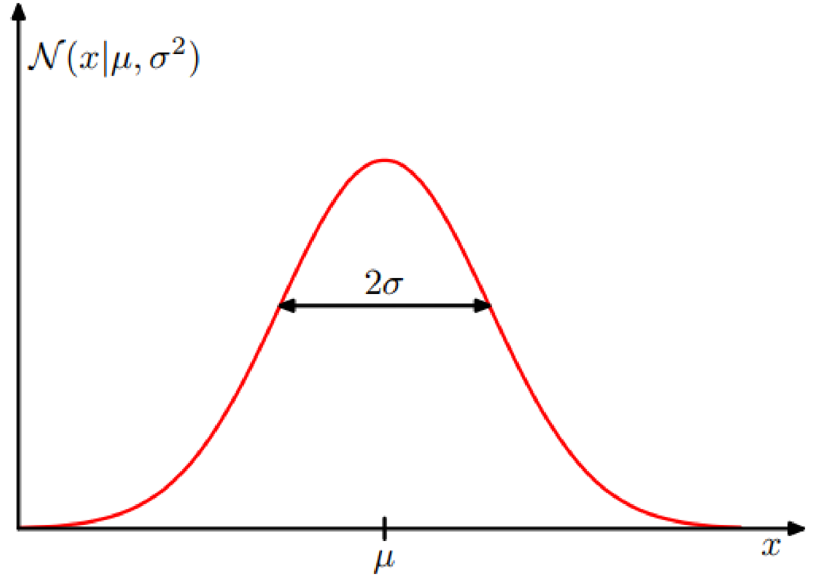
The Gaussian distribution

- The Gaussian or normal distribution is the most important distribution for continues variables.
- For the case of a single real-valued variable x, the Gaussian distribution is defined by

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

  - ❑ μ: mean
  - ❑ σ^2: variance
  - ❑ σ: standard deviation
  - ❑ $\frac{1}{\sigma^2}$: precision

- $\mathcal{N}(x|\mu,\sigma^2) \geq 0$
- $\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2) = 1$



$$p(\mu-\sigma \leq x \leq \mu+\sigma) \simeq 0.68$$

$$p(\mu-2\sigma \leq x \leq \mu+2\sigma) \simeq 0.95$$

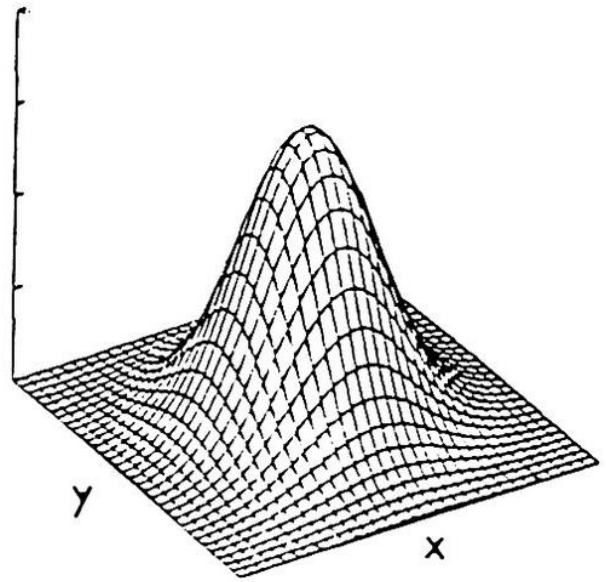
$$p(\mu-3\sigma \leq x \leq \mu+3\sigma) \simeq 0.99$$

The Gaussian distribution

- Gaussian distribution over a D-dimensional vector $\mathbf{x}$ of continuous variables

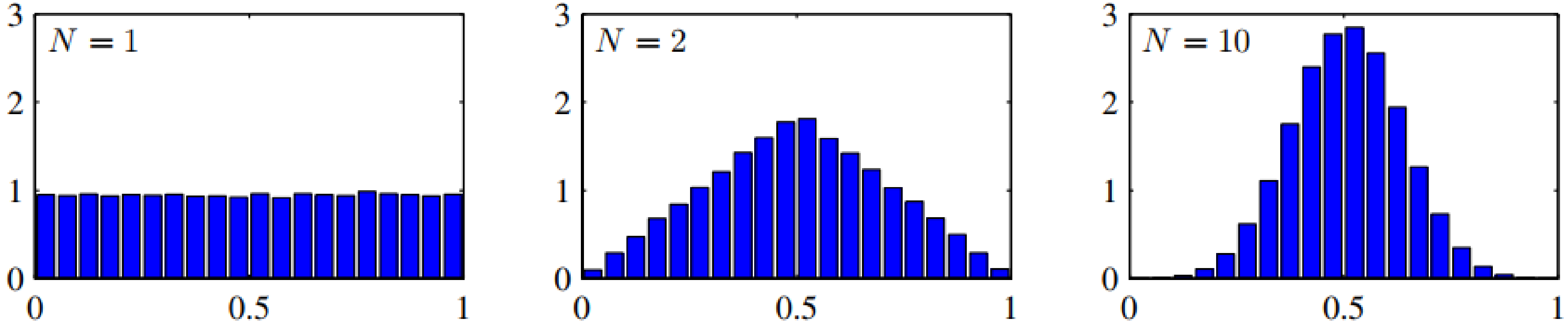
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

  - $\boldsymbol{\mu}$: a $D \times 1$ mean vector
  - $\boldsymbol{\Sigma}$: a $D \times D$ covariance matrix
  - $|\boldsymbol{\Sigma}|$: The determinant of $\boldsymbol{\Sigma}$



The Central Limit Theorem

- Mean of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases




```python
import numpy as np
from matplotlib import pyplot as plt
N = int(input())
means = []
for i in range(1,100000)
    means.append(np.mean(np.random.random(size=(N,))))

plt.hist(means,bins = 100, range=(0,1))
```

Prerequisites …

Eigenvectors and Eigenvalues

- For a square matrix $\mathbf{A}$ of size $M \times M$, the eigenvector equation is defined by

$$\mathbf{A}\mathbf{u}_i = \lambda_i \boldsymbol{u}_i \quad , i = 1, \dots, M$$

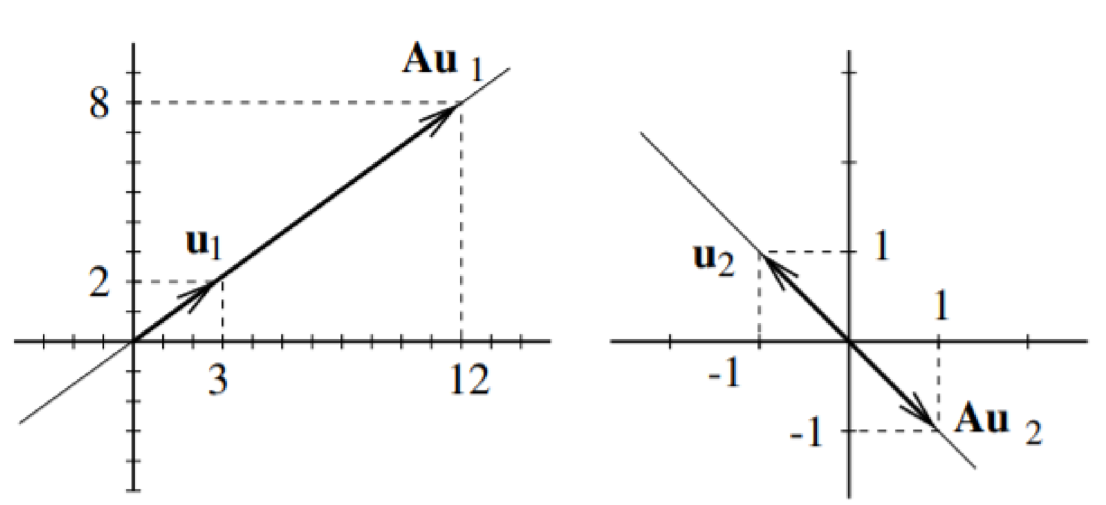
u_i: Eigenvector λ_i: Eigenvalue

- Characteristic Equation:

$$|\mathbf{A} - \lambda_i \mathbf{I}| = 0$$

- Example:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{cases} \mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \lambda_1 = 4 \\ \mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \lambda_1 = -1 \end{cases}$$



Prerequisites …

Eigenvectors and Eigenvalues

- For most applications we normalize the eigenvectors (i.e., transform them such that their length is equal to one)

$$\mathbf{u}_i \mathbf{u}_i^T = 1$$

- To normalize, we simply divide $\mathbf{u}_i$ by its length $|\mathbf{u}_i|$

- Example:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \lambda_1 = 4 \quad |\mathbf{u}_1| = \sqrt{3^2 + 2^2} = \sqrt{13}$$

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \lambda_1 = -1 \quad |\mathbf{u}_2| = \sqrt{-1^2 + 1^2} = \sqrt{2}$$

Normalized eigenvectors

$$\mathbf{u}_1 = \begin{bmatrix} 3/\sqrt{13} \\ 2/\sqrt{13} \end{bmatrix} = \begin{bmatrix} 0.8331 \\ 0.5547 \end{bmatrix}$$

$$\mathbf{u}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$$

Prerequisites …

Eigenvectors and Eigenvalues

- We can re-write the eigenvector equation in matrix form:

$$\mathbf{A}\mathbf{u}_\mathrm{i} = \lambda_i \boldsymbol{u}_i \xrightarrow{\text{Matrix Form}} \mathbf{AU} = \mathbf{U\Lambda}$$

Eigen Decomposition: $\mathbf{A} = \mathbf{U\Lambda U}^{-1}$

Diagonalization: $\mathbf{U}^{-1}\mathbf{AU} = \mathbf{\Lambda}$

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_M \end{pmatrix} \qquad \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_M \end{pmatrix}$$

$$\mathbf{A}\mathbf{u}_\mathrm{i} = \lambda_i \boldsymbol{u}_i \Rightarrow \mathbf{A}^{-1}\mathbf{A}\mathbf{u}_\mathrm{i} = \mathbf{A}^{-1}\lambda_i \boldsymbol{u}_i \Rightarrow \mathbf{u}_\mathrm{i} = \lambda_i \mathbf{A}^{-1}\mathbf{u}_i \Rightarrow \frac{1}{\lambda_i}\mathbf{u}_\mathrm{i} = \mathbf{A}^{-1}\mathbf{u}_i$$

Prerequisites …

Eigenvectors and Eigenvalues

- If $\mathbf{A}$ is a real symmetric matrix, then its eigenvalues are real and can be chosen to form orthonormal set, so that

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & , \text{if } i = j \\ 0 & , \text{otherwise} \end{cases}$$

Proof: Homework

- Or

$$\mathbf{U}^T\mathbf{U} = \mathbf{I} \quad \Rightarrow \quad \mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{U}^{-1} = \underbrace{\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}}_{☺}$$

- Then

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

A very nice property ☺

$$\mathbf{A}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\ \mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{M} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

Another nice property ☺

Prerequisites …

Eigenvectors and Eigenvalues

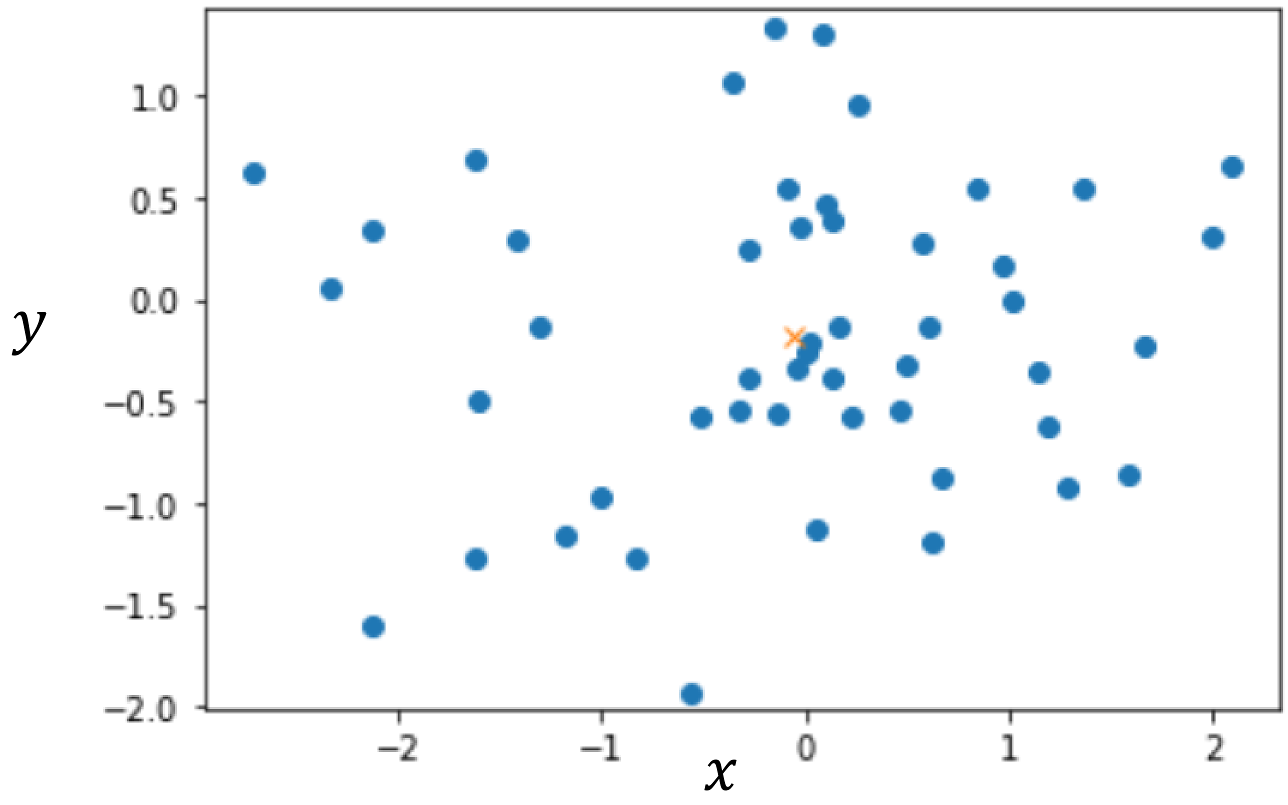
- The rank of matrix $\mathbf{A}$ is equal to the number of nonzero eigenvalues.
- A matrix $\mathbf{A}$ is called positive definite if its eigenvalues are strictly positive.
- A matrix $\mathbf{A}$ is called positive semidefinite if its eigenvalues are nonnegative.
- The product of the eigenvalues of $\mathbf{A}$ is the same as $|\mathbf{A}|$. Therefore, $\mathbf{A}$ is invertible if and only if it does not have a zero eigenvalue (its rank equals M)
- Generally the covariance matrix for the Gaussian distribution ($\boldsymbol{\Sigma}$) is symmetric and positive definite.

Prerequisites …

Mahalanobis Distance

- The Euclidean distance of a point from the mean (example for a 2D variable):

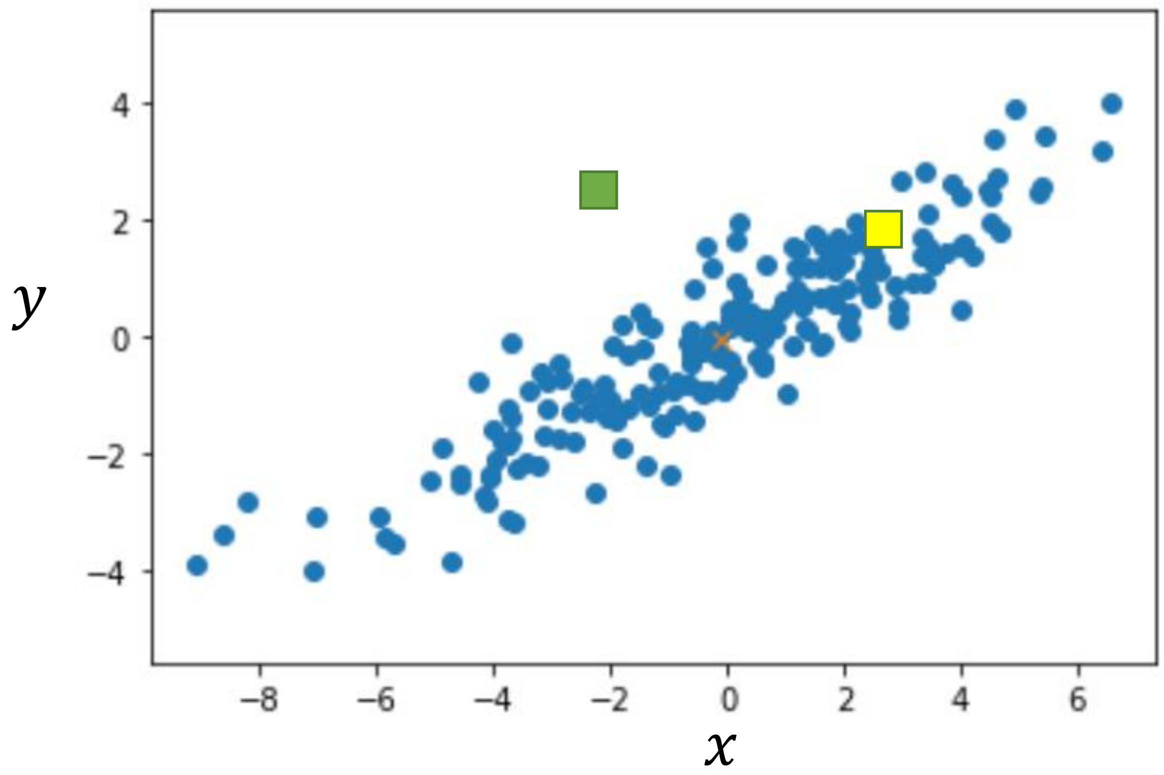
$$\sqrt{(x-\bar{x})^2+(y-\bar{y})^2}$$



Prerequisites …

Mahalanobis Distance

- However, Euclidean distance has limitations in real datasets, which often have some degree of covariance



Prerequisites …

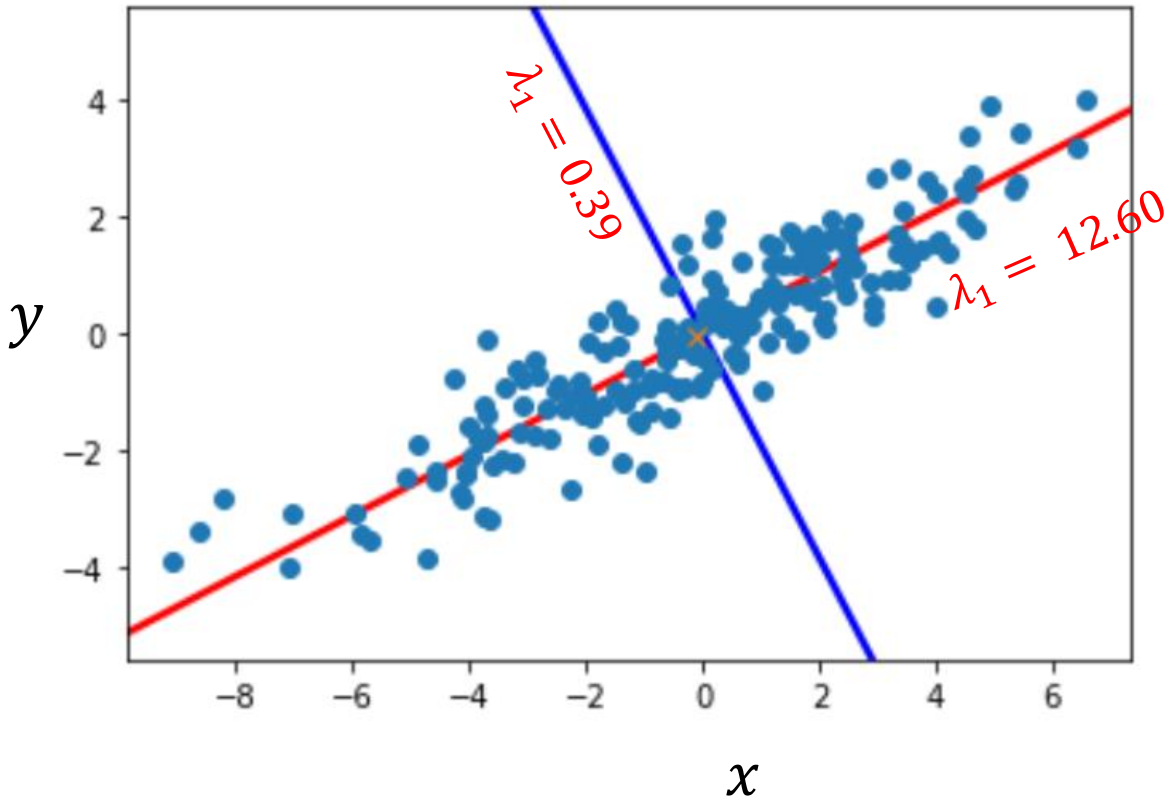
Mahalanobis Distance

- The idea of Mahalanobis distance is to remove the covariance by treating each eigenvector as a new axis, shrink the axis by $\sqrt{\lambda_i}$, then calculate distance between points

$$D^2 = (x - \bar{x})^T S^{-1} (x - \bar{x})$$

Data mean vector

Data covariance matrix



Prerequisites …

Jacobian Factor

- Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor.
- For instance, if we consider a change of variables $x = g(y)$, then a function $f(x)$ becomes $h(y) = f\big(g(y)\big)$
- Now consider a probability density $p_x(x)$
  - Observations falling in the range $(x, x + \delta x)$ have probability $p_x(x)\delta x$
  - By transforming them, we make them fall in the range $(y, y + \delta y)$
  - Observations falling in the range $(y, y + \delta y)$ have probability $p_y(y)\delta y$

$$p_x(x)\delta x = p_y(y)\delta y \Longrightarrow p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x(x)\left|\frac{dg(y)}{dy}\right| = p_x(x)|g'(y)|$$

- In the case of multivariate probabilities, in going from **x** to **y** coordinate system, we have:

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| \qquad \text{Where} \qquad \mathbf{J}_{\text{ij}} = \frac{\partial \mathbf{x}_i}{\partial \mathbf{y}_j}$$

The Gaussian distribution

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

- The geometric form of the Gaussian distribution
  - The Gaussian distribution depends on $\mathbf{x}$ is through the quadratic form

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

  - The quantity Δ is the Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$
  - This quantity reduces to the Euclidean distance when $\boldsymbol{\Sigma} = \mathbf{I}$
  - The Gaussian distribution will be constant on surfaces in x-space for which Δ^2 is constant.

The Gaussian distribution

- Consider the eigenvector equation for $\boldsymbol{\Sigma}$ (this matrix is real symmetric)

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad , \mathrm{i} = 1, \dots, D$$

- ❑ Eigenvectors form an orthonormal set

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & , \text{if } i = j \\ 0 & , \text{otherwise} \end{cases}$$

- ❑ $\boldsymbol{\Sigma}$ can be expressed as an expansion of its eigenvectors

$$\boldsymbol{\Sigma} = \sum_{i=1}^{\mathrm{D}} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

- ❑ The inverse covariance matrix can be expressed as

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{\mathrm{D}} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

The Gaussian distribution

- By substituting the inverse covariance matrix into the quadratic form Δ^2

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \quad = (\mathbf{x}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{\mathrm{D}}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)(\mathbf{x}-\boldsymbol{\mu}) \quad = \sum_{i=1}^{\mathrm{D}}\frac{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})}{\lambda_i}$$

$$= \sum_{i=1}^{\mathrm{D}}\frac{\mathrm{y}_i^2}{\lambda_i}$$

With $\mathrm{y}_i = \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$

- Forming the vector $\mathbf{y} = (y_1, \dots, y_D)^T$ we have:

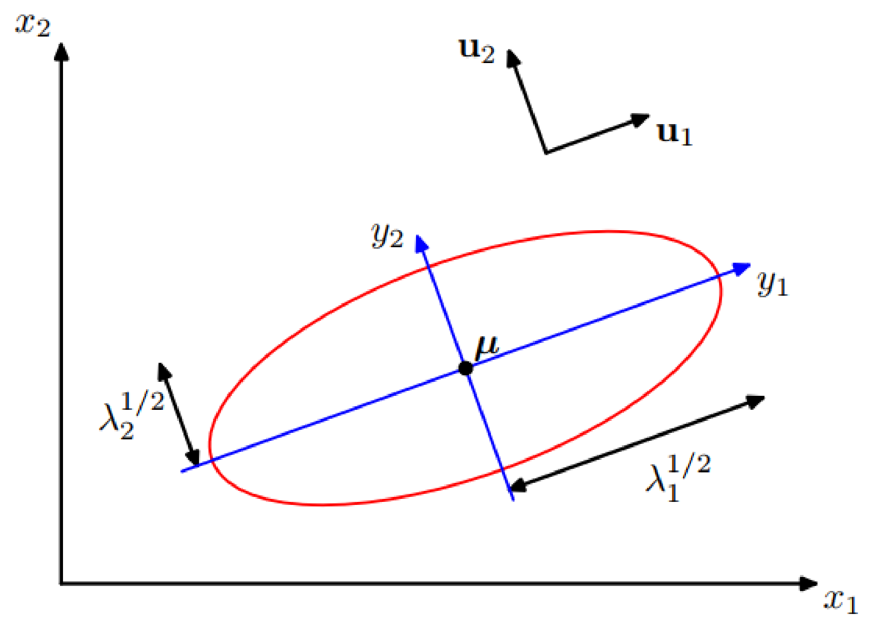
$$\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$$

- $\mathbf{U}$ is an orthogonal matrix whose rows are $\mathbf{u}_i^T$ (i.e., $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$)

The Gaussian distribution

- We can interpret $\{y_i\}$ as a new coordinate system defined by the orthonormal vectors $\mathbf{u}_i$ that are shifted and rotated with respect to the original x_i coordinates.

- The quadratic form, and hence the Gaussian density, will be constant on surfaces for which $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$ is constant.



- For positive λ_i, the surfaces are ellipsoids
  - ❑ Centered in $\boldsymbol{\mu}$ and axis oriented along $\mathbf{u}_i$.
  - ❑ The scaling factor in the directions of the axis are $\lambda_i^{\frac{1}{2}}$

The Gaussian distribution

- Now consider the form of the Gaussian distribution in the new coordinate system defined by the y_i.
- In going from the **x** to the **y** coordinate system, we have a Jacobian matrix **J**

$$\mathbf{J}_{\mathrm{ij}} = \frac{\partial \mathrm{x_i}}{\partial \mathrm{y_j}} = \mathbf{U}_{\mathrm{ij}}^{\mathrm{T}}$$

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu})$$

$$\Rightarrow \mathbf{x} = \mathbf{U}^{-\mathbf{1}}\mathbf{y} + \boldsymbol{\mu} = \mathbf{U}^{T}\mathbf{y} + \boldsymbol{\mu}$$

- Using the orthonormality property of the matrix **U**:

$$|\mathbf{J}|^2 = |\mathbf{U}^T|^2 = |\mathbf{U}^T||\mathbf{U}^T| = |\mathbf{U}^T||\mathbf{U}| = |\mathbf{U}^T\mathbf{U}| = |\mathbf{I}| = 1 \implies |\mathbf{J}| = 1$$

- Moreover

$$|\Sigma| = \prod_{j=1}^{D} \lambda_j \implies |\Sigma|^{\frac{1}{2}} = \prod_{j=1}^{D} \lambda_j^{\frac{1}{2}}$$

The Gaussian distribution

- Thus in the y_j coordinate system, the Gaussian distribution takes the form

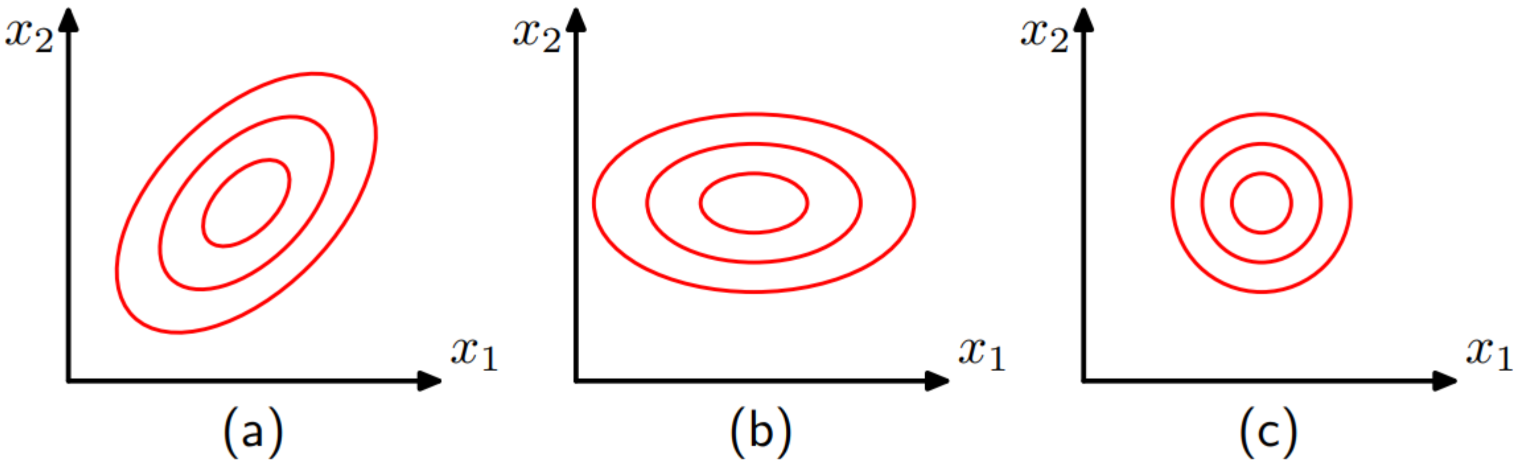
$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = p(\mathbf{x})$$

$$\Rightarrow p(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\Delta^2\right\} \xRightarrow{\Delta^2=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \; and \; |\Sigma|^{\frac{1}{2}}=\prod_{j=1}^{D}\lambda_j^{\frac{1}{2}}} p(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\prod_{j=1}^{D}\lambda_j^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}\right\}$$

$$= \frac{1}{\prod_{j=1}^{D}(2\pi)^{\frac{D}{2}}\lambda_j^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}\right\} = \prod_{j=1}^{D}\frac{1}{\left(2\pi\lambda_j\right)^{D/2}} \exp\left\{-\frac{y_i^2}{2\lambda_i}\right\}$$

- Therefore $p(\mathbf{y})$ is the product of D independent univariate Gaussian distributions

The Gaussian distribution



- ❑ a) General Σ ⟶ $\frac{D(D+3)}{2}$ parameters
- ❑ b) Diagonal Σ ⟶ $2D$ parameters
- ❑ c) $\Sigma = \sigma^2 \mathbf{I}$ ⟶ $D + 1$ parameters

Conditional Gaussian distribution

Property 1 of the Gaussian Distribution: If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.

$$p(\mathbf{x},\mathbf{y}) = \mathcal{N}\left((\mathbf{x},\mathbf{y})|\boldsymbol{\mu}_{(\mathbf{x},\mathbf{y})},\boldsymbol{\Sigma}_{(\mathbf{x},\mathbf{y})}\right) \Longrightarrow p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left((\mathbf{x}|\mathbf{y})|\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}},\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$$

- Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$
  - we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$.
  - Without loss of generality, we can take $\mathbf{x}_a$ to form the first M components of $\mathbf{x}$, with $\mathbf{x}_b$ comprising the remaining $\mathrm{D}-\mathrm{M}$ components,
  - Then

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

Conditional Gaussian distribution

- We also define corresponding partitions of
  - the mean vector $\boldsymbol{\mu}$ given by $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$
  - The covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$
  - Because $\boldsymbol{\Sigma}$ is symmetric ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}}$) then $\boldsymbol{\Sigma}_{aa}$ and $\boldsymbol{\Sigma}_{bb}$ are also symmetric and $\boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^T$

- In many situations, it is convenient to work with the precision matrix: $\Lambda = \Sigma^{-1}$
  - The corresponding partition for $\boldsymbol{\Lambda}$: $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$
  - Because the inverse of a symmetric matrix is also symmetric then $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{bb}$ are also symmetric and $\boldsymbol{\Lambda}_{ab} = \boldsymbol{\Lambda}_{ba}^T$
  - It should be stressed that, for instance, $\boldsymbol{\Lambda}_{aa}$ is not simply given by the inverse of $\boldsymbol{\Sigma}_{aa}$.

Conditional Gaussian distribution

- We have

$$p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\Rightarrow p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_b)} = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{\int_{-\infty}^{+\infty} p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_a}$$

Not an easy approach ☹

Better approach (Analytical Method):

- Remember
  - The Gaussian distribution depends on **x** is through the quadratic form

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Therefore, to show that $p(\mathbf{x}_a|\mathbf{x}_b)$ is Gaussian, we need to proof that $p(\mathbf{x}_a|\mathbf{x}_b)$ has a similar quadratic form with respect to $\mathbf{x}_a$.

Conditional Gaussian distribution

- We have

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

$$= \left(\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \right)^T \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \left(\begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \right)$$

$$= -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa} (\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{\boldsymbol{ab}} (\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})$$
$$-\frac{1}{2}(\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})^T \boldsymbol{\Lambda}_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{\boldsymbol{bb}} (\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})$$

- We see that as a function of $\mathbf{x}_a$, this is a quadratic form, and hence the corresponding conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ will be Gaussian.

Conditional Gaussian distribution

- **Question**: How to find $\boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b}$ for $p(\mathbf{x}_a|\mathbf{x}_b)$?
- **Answer**: Using an approach called Completing the Square
  - For a general Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the exponent can be written as:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathrm{const}$$

  - For $p(\mathbf{x}_a|\mathbf{x}_b)$, we have:

$$\Delta^2 = -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{\boldsymbol{ab}}(\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})$$
$$-\frac{1}{2}(\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})^T \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{\boldsymbol{bb}}(\mathbf{x}_{\boldsymbol{b}} - \boldsymbol{\mu}_{\boldsymbol{b}})$$

  - If we pick out all terms that are second order in $\mathbf{x}_a$, we have

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a \quad \Rightarrow \quad \boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b}^{-1} = \boldsymbol{\Lambda}_{aa} \Rightarrow \boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\Lambda}_{aa}^{-1}$$

Conditional Gaussian distribution

- For $p(\mathbf{x}_a|\mathbf{x}_b)$, we have:

$$\Delta^2 = -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$
$$-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

- If we pick out all terms that are linear in $\mathbf{x}_a$, we have

$$\mathbf{x}_a^{\mathrm{T}}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} = \mathbf{x}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b}^{-1}\boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b}$$

$\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\Lambda}_{aa}^{-1}$ ⟹

$$\boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\}$$
$$= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

Conditional Gaussian distribution

- **Summary:** $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$ $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$ $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$ $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$

$$\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\Lambda}_{aa}^{-1} \qquad \boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

- **Question**: How to find $\boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b}$ in terms of $\boldsymbol{\Sigma}$ (not $\boldsymbol{\Lambda}$)

  - We can use the following identity:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \quad \text{Where} \quad \mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$$

  - $$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \Rightarrow \begin{aligned} \boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \end{aligned}$$

  - $$\boldsymbol{\Sigma}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\Lambda}_{aa}^{-1} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba}) \qquad \boldsymbol{\mu}_{\mathbf{x}_a|\mathbf{x}_b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

Marginal Gaussian distribution

Property 2 of the Gaussian Distribution: If two sets of variables are jointly Gaussian, then the marginal distributions is again Gaussian.

$$p(\mathbf{x},\mathbf{y}) = \mathcal{N}\left((\mathbf{x},\mathbf{y})|\boldsymbol{\mu}_{(\mathbf{x},\mathbf{y})},\boldsymbol{\Sigma}_{(\mathbf{x},\mathbf{y})}\right) \Longrightarrow p(\mathbf{x}) = \mathcal{N}\left((\mathbf{x})|\boldsymbol{\mu}_{\mathbf{x}},\boldsymbol{\Sigma}_{\mathbf{x}}\right)$$

- $$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$
- Similar to conditional probability, we can prove that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad \Longrightarrow \quad p(\mathbf{x}_{\mathrm{a}}) = \mathcal{N}(\mathbf{x}_{\mathrm{a}}|\boldsymbol{\mu}_{\mathrm{a}},\boldsymbol{\Sigma}_{\mathrm{aa}})$$

Bayes' Theorem for Gaussian Variables

Property 3 of the Gaussian Distribution: Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

The joint distribution of $\mathbf{x}$ and $\mathbf{y}$ is given by

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \middle| \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix}\right)$$

- We find an expression for the joint distribution $p(\mathbf{x}, \mathbf{y})$.
  - To do this, we define

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

Bayes' Theorem for Gaussian Variables

- Then we have

$$p(\mathbf{z}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \times \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

- Considering the log of the joint distribution

$$\ln p(\mathbf{z}) = \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x})$$

$$= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) + \text{const}$$

- As before, we see that this is a quadratic function of the components of $\mathbf{z}$, and hence $p(\mathbf{z})$ is Gaussian distribution.

- To find the precision of this Gaussian, we consider the second order terms

Remember: Completing the Squares

$$(\mathbf{x} - \boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

$$-\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$

$$= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{z} \Rightarrow \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}$$

Bayes' Theorem for Gaussian Variables

Remember $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix}$ Where $\mathbf{M} = \left(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}\right)^{-1}$

- Then

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{LA} & -\mathbf{A}^{\mathrm{T}}\mathbf{L} \\ -\mathbf{LA} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix}$$

- Similarly, we can find the mean of the Gaussian distribution over z by identifying the linear terms

Remember: Completing the Squares

$$(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{Lb} + \mathbf{y}^{\mathrm{T}}\mathbf{Lb} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{Lb} \\ \mathbf{Lb} \end{pmatrix} = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} \Rightarrow$$

$$\boldsymbol{\mu}_{\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{z}} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{Lb} \\ \mathbf{Lb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{Lb} \\ \mathbf{Lb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}$$

Bayes' Theorem for Gaussian Variables

Property 4 of the Gaussian Distribution: Given a marginal Gaussian distribution for **x** and a conditional Gaussian distribution for **y** given **x** in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

The marginal distribution of y is given by

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)$$

- It is obvious from properties 2 and 3.

Remember

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longrightarrow p(\mathbf{x}_\mathrm{a}) = \mathcal{N}(\mathbf{x}_\mathrm{a}|\boldsymbol{\mu}_\mathrm{a}, \boldsymbol{\Sigma}_\mathrm{aa})$$

Bayes' Theorem for Gaussian Variables

Property 5 of the Gaussian Distribution: Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

The conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}\right) \quad \text{Where} \quad \boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A}\right)^{-1}$$

- It is obvious from properties 1 and 3.

Remember

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longrightarrow p(\mathbf{x}_\mathrm{a}|\mathbf{x}_\mathbf{b}) = \mathcal{N}(\mathbf{x}_\mathrm{a}|\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b), \boldsymbol{\Lambda}_{aa}^{-1})$$

Maximum Likelihood for the Gaussian

- Given a data set $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^{\mathrm{T}}$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood.

$$\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \prod_{n=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x_i}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_i}-\boldsymbol{\mu})\right\}$$

$$= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \ln \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x_i}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_i}-\boldsymbol{\mu})\right\}$$

$$= \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} -\frac{\mathrm{ND}}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{ML})(\mathbf{x}_n-\boldsymbol{\mu}_{ML})^{\mathrm{T}}$$

Proof: Homework

Maximum Likelihood for the Gaussian

- In the case of $D = 1$ (univariate Gaussian distribution):

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

- **Question**: Are $\boldsymbol{\mu}_{ML}$ and $\boldsymbol{\Sigma}_{ML}$ good estimations for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$?
  - A good estimation $\hat{\alpha}$ for parameter α should be unbiased to the data set

$$\mathbb{E}[\hat{\alpha}] = \alpha$$

  - For the cases of $\boldsymbol{\mu}_{ML}$ and $\boldsymbol{\Sigma}_{ML}$ we have:

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu} \qquad \mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma}$$

☺ ☹

- A better (unbiased) estimation for σ^2

$$\widetilde{\boldsymbol{\Sigma}} = \frac{N}{N-1}\boldsymbol{\Sigma}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^{\mathrm{T}}$$

Bayesian Inference for the Gaussian

- The maximum likelihood framework gave point estimates for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Now we develop a Bayesian treatment by introducing prior distributions over these parameters.

- We will consider the following cases
  - The variance is known, and we consider the task of inferring the mean
  - The mean is known, and we consider the task of inferring the variance

Bayesian Inference for the Gaussian

- **Case 1:** The variance is known, and we consider the task of inferring the mean
  - Let us begin with a simple example in which we consider a single Gaussian random variable x (D = 1) .
  - We shall suppose that the variance σ^2 is known, and we consider the task of inferring the mean μ given a set of N observations $\mathbf{X} = \{x_1, \dots, x_N\}$.
  - The likelihood function:

Remember $p(\mu|\mathcal{D}) = \dfrac{p(\mathcal{D}|\mu) \times p(\mu)}{p(D)}$

$$p(\mathbf{X}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

  - The likelihood function takes the form of the exponential of a quadratic form in μ. If we choose a Gaussian prior $p(\mu)$, it will be a conjugate distribution for the likelihood function.
  - The posterior is a product of two exponentials of quadratic functions of μ and hence will also be Gaussian.

Bayesian Inference for the Gaussian

- We take our prior distribution to be

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

- The posterior distribution is given by

$$p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$$

- Some manipulations involving completing the square in the exponent allow to show that the posterior distribution is given by

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Bayesian Inference for the Gaussian

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \qquad \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

- Note that μ_N (the mean of the posterior distribution) is a compromise between the prior mean (μ_0) and the maximum likelihood solution μ_{ML}.
  - If $N = 0$, μ_N reduces to the prior mean.
  - For $N \to \infty$, the posterior mean equals the maximum likelihood solution.

- As we increase the number of observed data points, the precision steadily increases, corresponding to a posterior distribution with steadily decreasing variance.
  - With no observed data points, we have the prior variance
  - If $N \to \infty$, the variance $\sigma_N^2 \to 0$ and the posterior distribution becomes infinitely peaked around μ_{ML}

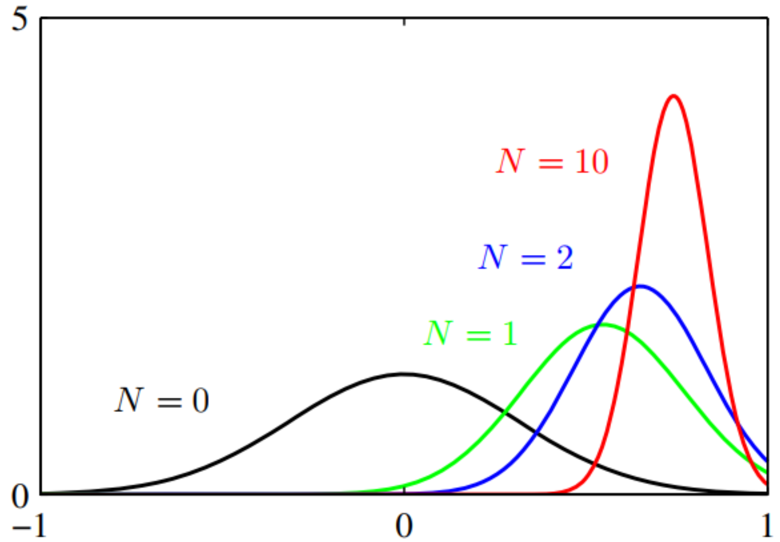


Figure. The data points are generated from a Gaussian of mean 0.8 and variance 0.1, and the prior is chosen to have mean 0.

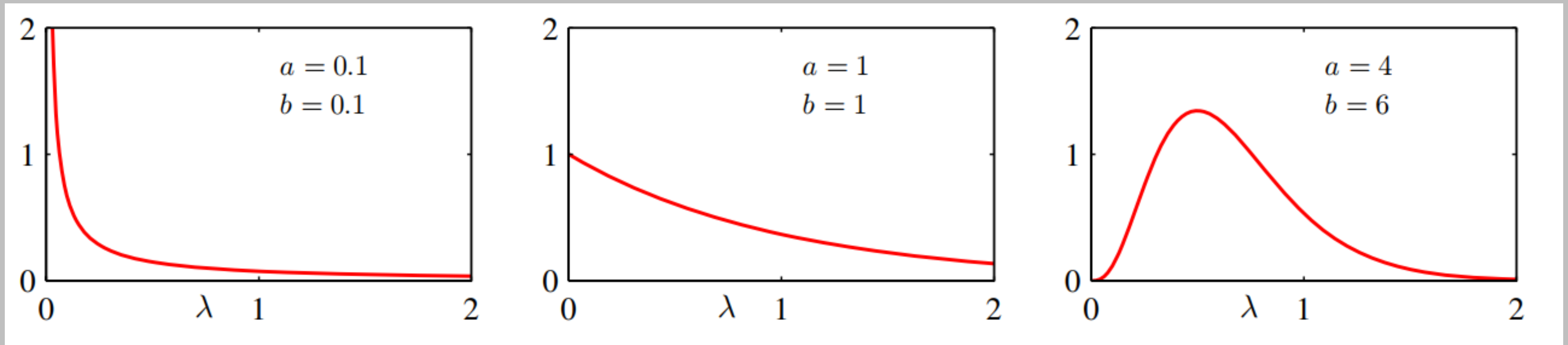
Bayesian Inference for the Gaussian

- **Case 2:** The mean is known, and we consider the task of inferring the variance
  - The likelihood function (It turns out to be most convenient to work with the precision $\lambda \equiv \frac{1}{\sigma^2}$):

$$p(\mathbf{X}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{\frac{N}{2}} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

  - The corresponding conjugate prior should therefore be proportional to the product of a power of λ and the exponential of a linear function of λ.
  - This corresponds to the gamma distribution which is defined by

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \qquad \mathbb{E}[\lambda] = \frac{a}{b}, \; var[\lambda] = \frac{a}{b^2}$$



Bayesian Inference for the Gaussian

- Consider a prior distribution $\mathrm{Gam}(\lambda|a_0, b_0)$. Multiplying by the likelihood function, we obtain a the following posterior distribution

$$p(\lambda|\mathbf{X}) \propto \lambda^{a_0-1}\lambda^{\frac{N}{2}} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- which we recognize as a gamma distribution of the form $\mathrm{Gam}(\lambda|a_N, b_N)$ where

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = \mathrm{b}_0 + \frac{\mathrm{N}}{2}\sigma_{ML}^2$$

Mixtures of Gaussians

- While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modelling real data sets.

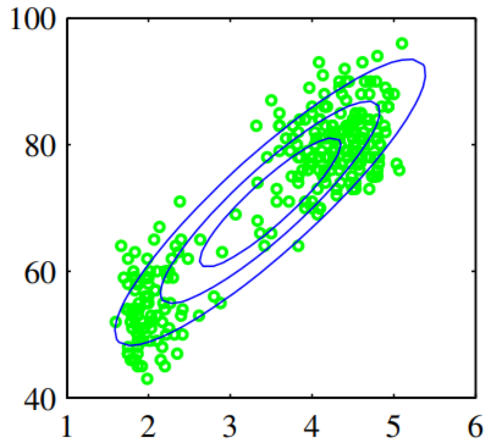


Figure: A single Gaussian distribution fitted to the data using maximum likelihood.

- Note that this distribution fails to capture the two clumps in the data and indeed places much of its probability mass in the central region between the clumps where the data are relatively sparse.

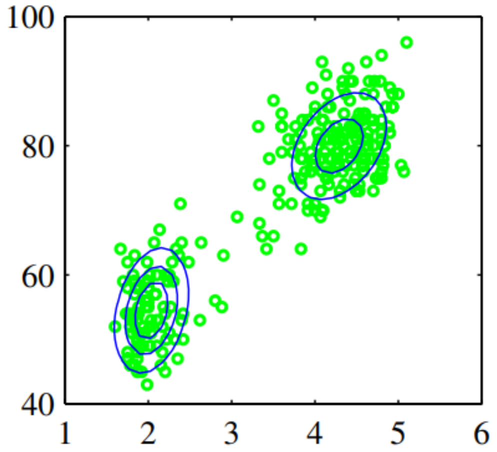


Figure: A linear combination of two Gaussians fitted using maximum likelihood

- Such superpositions, formed by taking linear combinations of more basic distributions such as Gaussians, can be formulated as probabilistic models known as mixture distributions

Mixtures of Gaussians

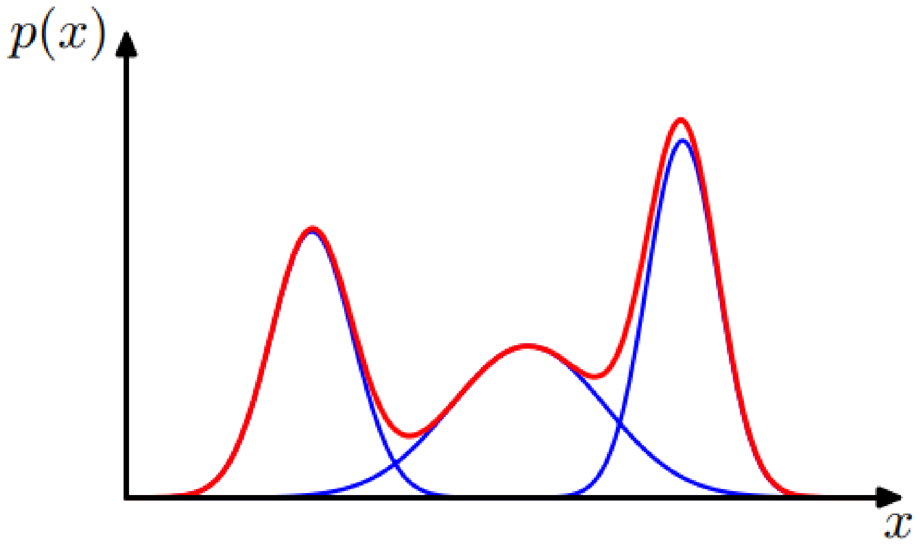


Figure: Example of a Gaussian mixture distribution in one dimension

- Three Gaussians (blue) and their sum (red)
- We can get very complex densities

- By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy.
- We consider a mixture of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Mixtures of Gaussians

- We consider a mixture of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  - ❑ The parameters π_k are called mixing coefficients. $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

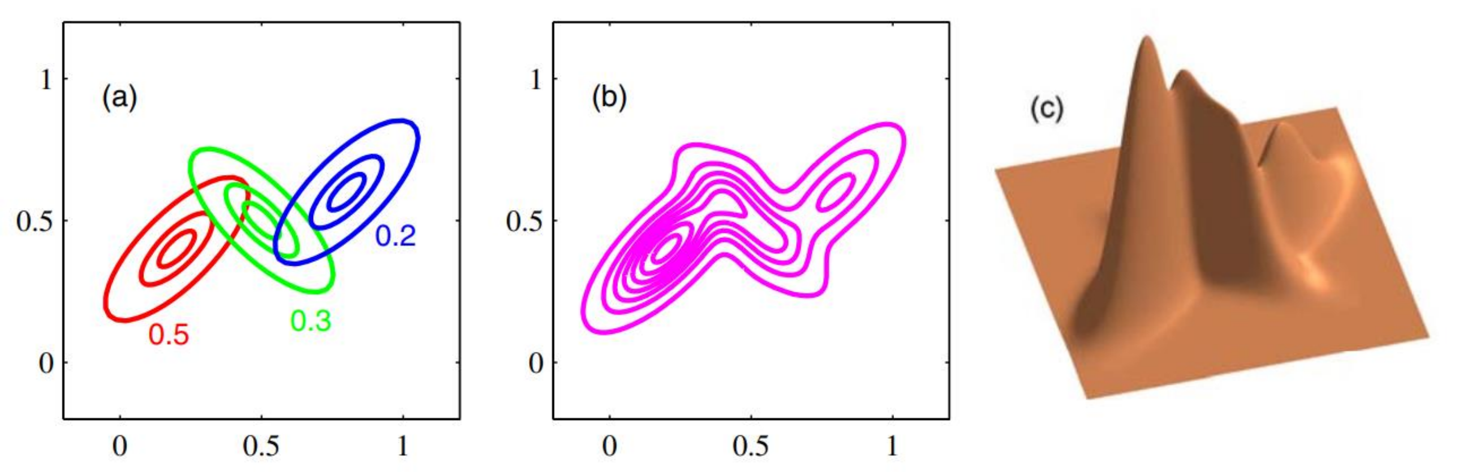


Figure: A mixture of 3 Gaussians in a two-dimensional space

- (a) Contours of constant density for each of the mixture components (b) Contours of the mixture distribution $p(\mathbf{x})$. (c) A surface plot of the distribution $p(\mathbf{x})$.