Probability Distributions 1

수학과 오서영

1. Binary Variables

Single binary random variable $x \in \{0, 1\}$
example) x : outcome of flipping a coin (x=1 : head, x=0 : tail)
Damaged coin? -> probability of landing heads is not same

The probability of x=1 : $p(x = 1|\mu) = \mu$ where $0 \leqslant \mu \leqslant 1$

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

Bernoulli distribution :

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

$$\begin{aligned} \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1 - \mu) \end{aligned}$$

1. Binary Variables

Dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ (observed values)

Likelihood function (function of μ)

- Assumption : observations are independent from $p(x|\mu)$

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

We can estimate μ by **maximizing** the likelihood function

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

-> $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$ (sample mean)

proof

$$\frac{\partial}{\partial \mu}\ln p(\mathcal{D}|\mu) = 0 \Leftrightarrow$$

$$\frac{\partial}{\partial \mu}\sum_{n=1}^{N}\Big(x_n \ln \mu + (1-x_n)\ln(1-\mu)\Big) = 0 \Leftrightarrow$$

$$\sum_{n=1}^{N}\frac{\partial}{\partial \mu}\Big(x_n \ln \mu + (1-x_n)\ln(1-\mu)\Big) = 0 \Leftrightarrow$$

$$\sum_{n=1}^{N}\left(\frac{1}{\mu}x_n - \frac{1}{1-\mu}(1-x_n)\right) = 0 \Leftrightarrow$$

$$\sum_{n=1}^{N}\left(\frac{1}{\mu}x_n - \frac{1}{1-\mu} + \frac{1}{1-\mu}x_n\right) = 0 \Leftrightarrow$$

$$\sum_{n=1}^{N}\left(\frac{1}{\mu}x_n + \frac{1}{1-\mu}x_n\right) = \frac{N}{1-\mu} \Leftrightarrow$$

$$\sum_{n=1}^{N}\left(\frac{1-\mu}{\mu}x_n + x_n\right) = N \Leftrightarrow$$

$$\sum_{n=1}^{N}\frac{1}{\mu}x_n = N \Leftrightarrow$$

1. Binary Variables

m : The number of observations of x=1

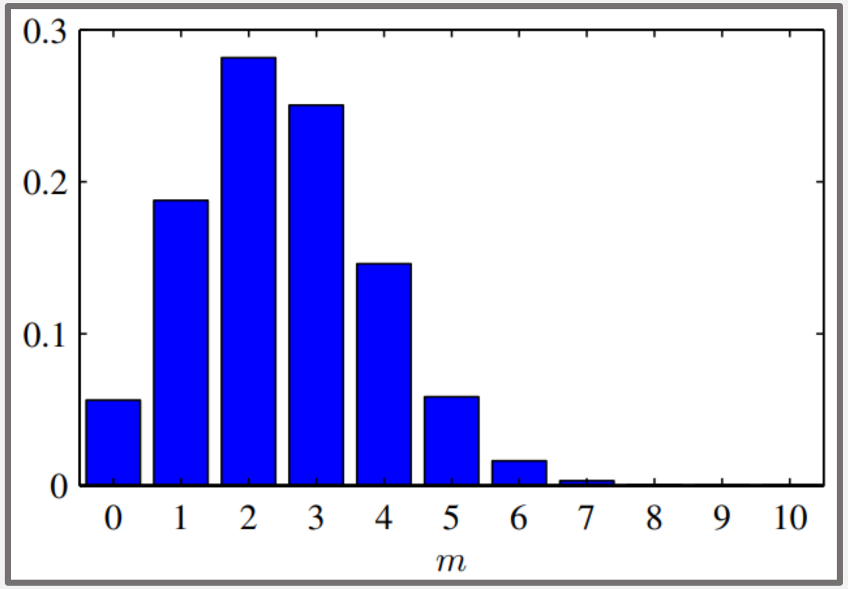
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

Overfitting by observed dataset -> prior distribution?
Binomial distribution :

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

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

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu) = N\mu(1-\mu).$$



Binomial distribution with
N=10, mu = 0.25

2.1.1 The beta distribution

Binomial distribution : overfitting for small dataset
-> prior : proportional to powers of μ and 1- μ

Beta distribution :

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

normalization

$$\int_0^1 \mathrm{Beta}(\mu|a,b)\,\mathrm{d}\mu = 1.$$

where $\Gamma(x)$ is the gamma function

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

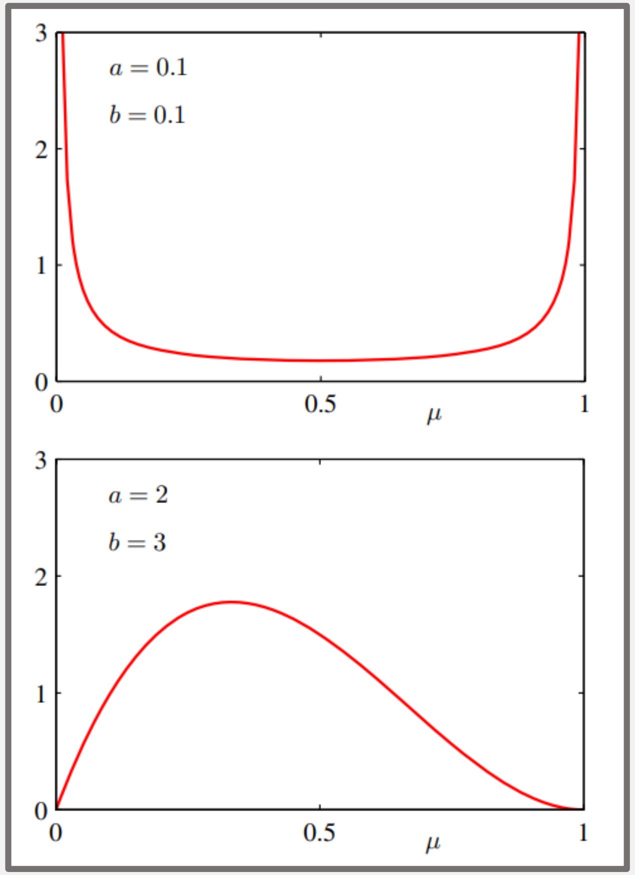
$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}.$$

2.1.1 The beta distribution

Posterior distribution over μ.
: multiplying beta prior by the binomial likelihood function and normalizing

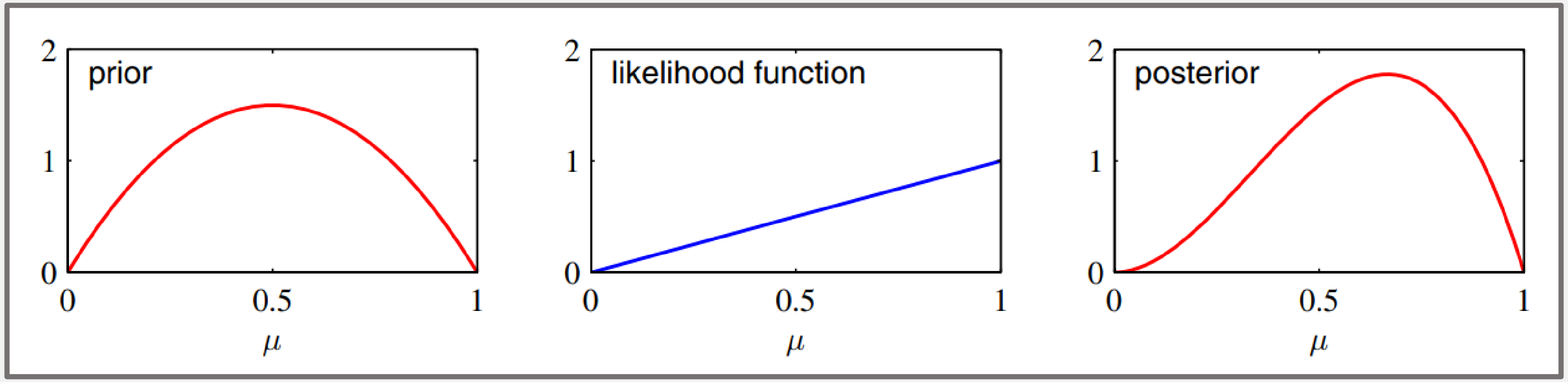
$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}.$$

where $l = N - m.$ (tail)
-> The posterior distribution also follows the beta distribution.
a, b : effective number of observation



a, b : hyperparameters

2.1.1 The beta distribution



Sequential Bayesian inference

Prior : beta with a=2, b=2 **Likelihood** : N = 1 **Posterior** : beta with a=3, b=2

2.1.1 The beta distribution

Sequential approach to learning
Maximum likelihood methods can also be cast into a sequential framework

- the posterior distribution can act as the prior if subsequent observations arrive

-> taking one observation at a time and after each observation **updating** the current posterior distribution by multiplying by the likelihood of the incoming observation

- It is **independent** of the choice of prior and of the likelihood function and **depends only** on the assumption of data

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})\,\mathrm{d}\mu = \int_0^1 \mu p(\mu|\mathcal{D})\,\mathrm{d}\mu = \mathbb{E}[\mu|\mathcal{D}].$$

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$$

- If m, l are large, MLE!

2.2. Multinomial Variables

Discrete random variables $\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}$
where k is represented k-th element being 1 and all others being 0.
-> $x_k = 1$ by the parameter μ_k

Generalization of Bernoulli distribution -> Categorical distribution :

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$ where $\mu_k \geqslant 0$ and $\sum_k \mu_k = 1$

Since parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ (vector),

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu}.$$

proof

$$\begin{aligned}\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) &= \sum_{\mathbf{x}} \prod_{k=1}^{K} \mu_k^{x_k} \\ &= \prod_{k=1}^{K} \mu_k^{x_k^1} + \cdots + \prod_{k=1}^{K} \mu_k^{x_k^K} \\ &= \sum_{k=1}^{K} \mu_k = 1\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] \\ &= \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}) \\ &= \sum_{\mathbf{x}} \mathbf{x} \prod_{k=1}^{K} \mu_k^{x_k} \\ &= \mathbf{x}^1 \prod_{k=1}^{K} \mu_k^{x_k^1} + \cdots + \mathbf{x}^K \prod_{k=1}^{K} \mu_k^{x_k^K} \\ &= (\mu_1, \ldots, \mu_K)^{\mathrm{T}}\end{aligned}$$

2.2. Multinomial Variables

dataset D of N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$
likelihood function :

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

-> likelihood function depends on N data only through the K quantities :

$$m_k = \sum_n x_{nk}$$

-> **Sufficient statistics**

2.2.1 The Dirichlet distribution

Family of prior distributions for parameters $\{\mu_k\}$ of multinomial distributions.

Conjugate prior : $p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$ where $0 \leqslant \mu_k \leqslant 1$

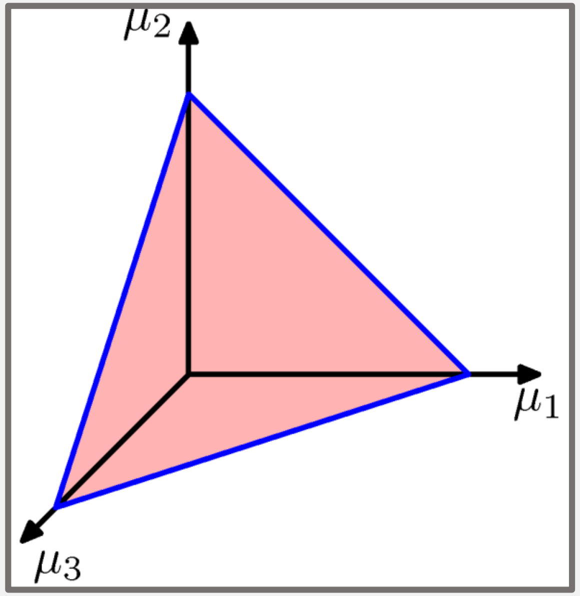
$\sum_k \mu_k = 1$

Dirichlet Distribution :

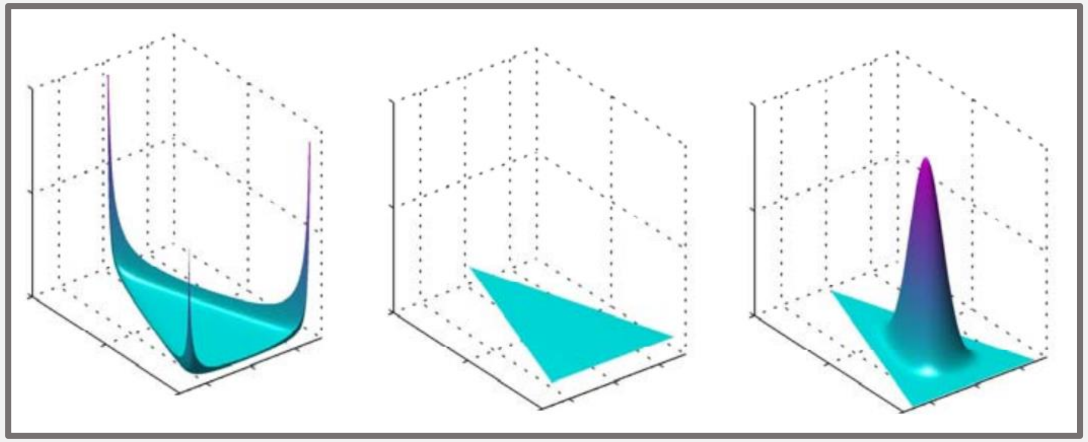
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

where $\Gamma(x)$: gamma function and $\alpha_0 = \sum_{k=1}^{K} \alpha_k.$

Simplex?



K-1 dimensions



2.3. The Gaussian Distribution

Continuous random variable x
Gaussian = normal distribution :

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

Multi-dimension

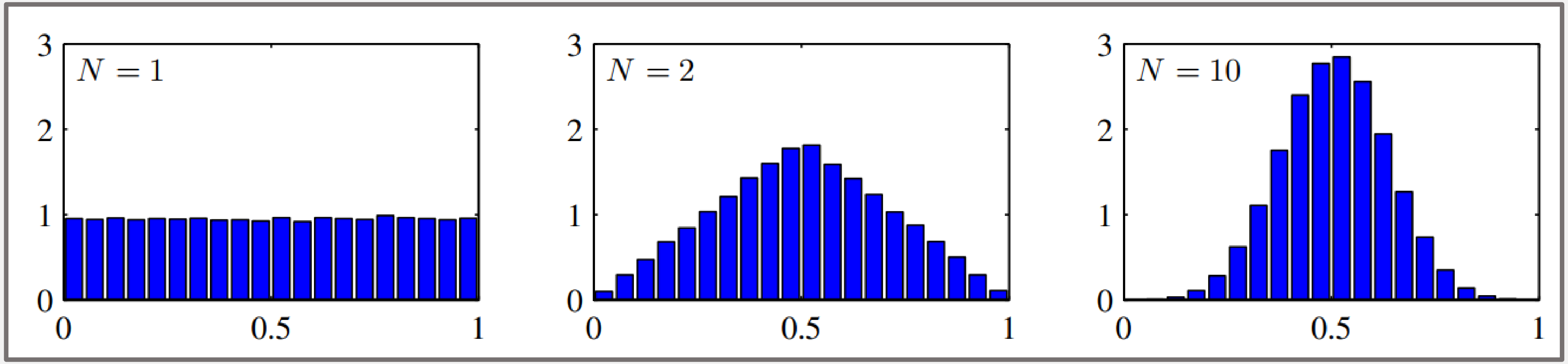
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where D by D covariance matrix

-> the distribution **maximizing the entropy** is a Gaussian distribution.

2.3. The Gaussian Distribution

Central limit theorem
The sum of a set of random variables has a distribution that becomes increasingly Gaussian as the number of terms increase.



2.3. The Gaussian Distribution

Important analytical properties of Gaussian Distribution

functional dependence of the Gaussian on x is through the quadratic form

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$ where Δ is Mahalanobis distance

- Eigenvector equation for covariance matrix : $\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i$

symmetric **orthonormal**

-> $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij}$

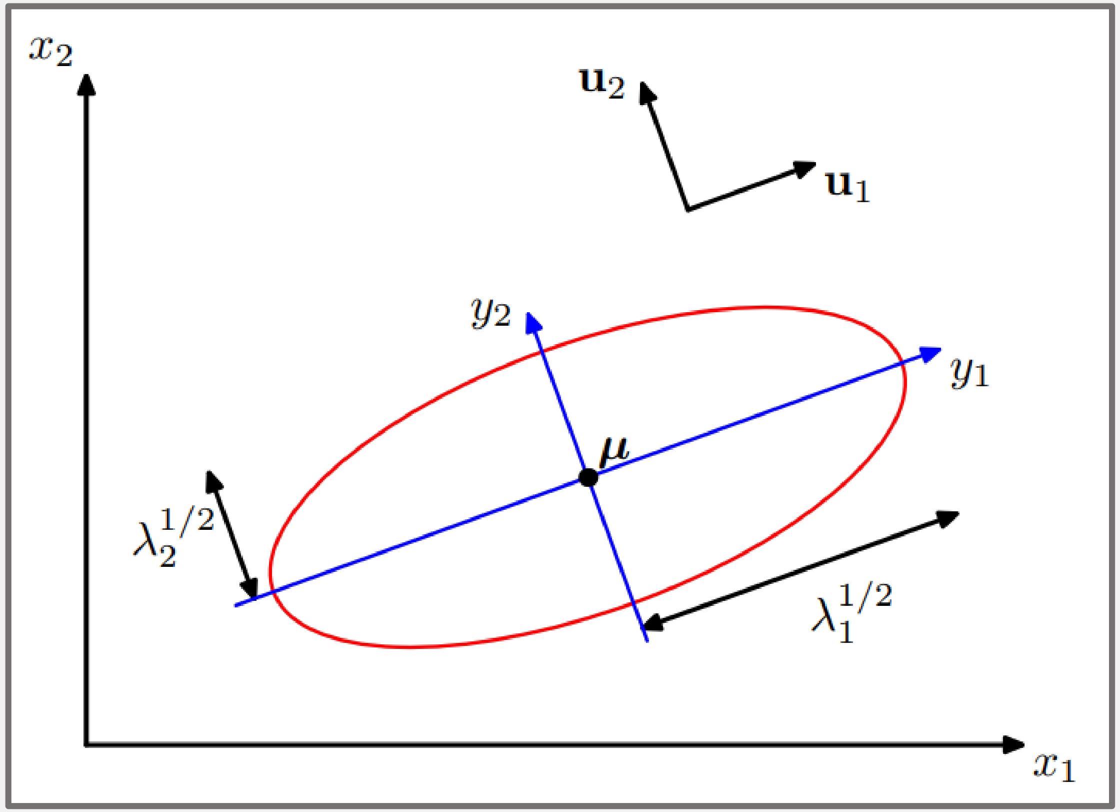
$$I_{ij} = \begin{cases} 1, & \text{if } i=j \\ 0, & \text{otherwise.} \end{cases}$$

-> $\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$ -> $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}. \qquad \mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu})$$

-> **y** : new coordinate system defined by the orthonormal vecter U

2.3. The Gaussian Distribution



Red
: elliptical surface of constant probability density of Gaussian

The major axes of the ellipse are defined by the eigenvectors U of covariance matrix

$\lambda_2^{1/2}$ scale ellipse, centroid : mu
-> i.e all eigenvalues > 0

2.3. The Gaussian Distribution

Gaussian distribution in the new coordinate system defined by y

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}$$ where J : jacobian matrix , U : orthonormal

-> $|\mathbf{J}|^2 = \left|\mathbf{U}^{\mathrm{T}}\right|^2 = \left|\mathbf{U}^{\mathrm{T}}\right| |\mathbf{U}| = \left|\mathbf{U}^{\mathrm{T}}\mathbf{U}\right| = |\mathbf{I}| = 1$ -> $|\mathbf{J}| = 1$

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}.$$

$\xrightarrow{\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu})}$ $$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

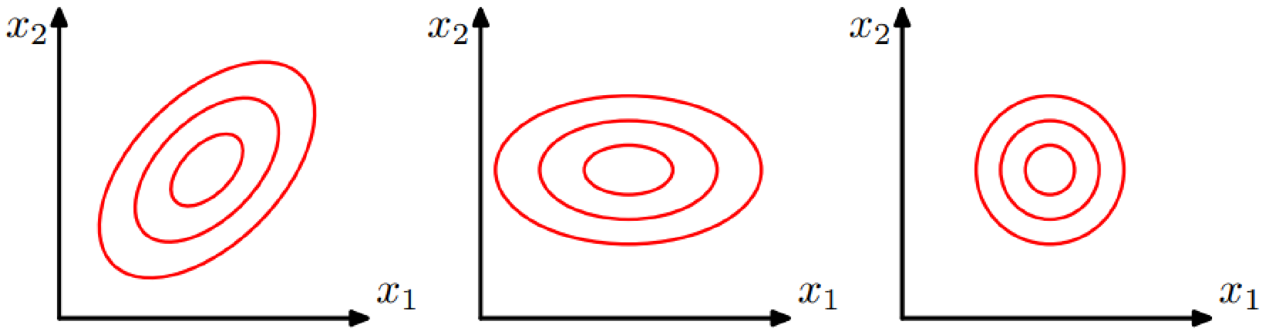
-> Product of independent normal distributions

2.3. The Gaussian Distribution

Significant limitations

1. General symmetric covariance matrix has D(D+1)/2 independent parameters and There are another independent parameters in mu -> total : D(D+3)/2 (many..)

-> restrict the covariance matrix to diagonal matrix or identity matrix



(a) **General form** (b) **diagonal** (c) **identity**

2. Gaussian distribution is intrinsically unimodal (i.e., has a single maximum)
-> unable to approximate multimodal distributions.

2.3.1 Conditional Gaussian distributions

x : D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Partition $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$ $\xrightarrow{\text{correspond}}$ partition of mean $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$

Covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$

Since $\boldsymbol{\Sigma}$ is symmetric, $\boldsymbol{\Sigma}^{\mathrm{T}} = \boldsymbol{\Sigma}$ -> $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^{\mathrm{T}}$.

Let $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ then $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$: precision matrix

$p(\mathbf{x}_a|\mathbf{x}_b)$? -> consider $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$

fix

2.3.1 Conditional Gaussian distributions

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

Quadratic form of $\mathbf{x}_a$ i.e $p(\mathbf{x}_a|\mathbf{x}_b)$: Gaussian (Covariance?)

Diff twice $\longrightarrow$ $-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a$ -> $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}.$

- Summary) $p(\mathbf{x}_a, \mathbf{x}_b)$: Gaussian -> $p(\mathbf{x}_a|\mathbf{x}_b)$: Gaussian