Chapter 2 Probability distributions

Introduction to useful probability distributions $\left\langle \begin{array}{l}\text{discrete}\\ \text{continuous}\end{array}\right.$

Discuss key statistical concepts such as Bayesian inference

Density estimation (find $p(x)$) independent and identically

- iid assumption observations $x_1 \dots x_N$
- unsupervised learning
- parametric vs nonparametric

Conjugate distributions (prior and posterior)

2.1 Discrete random variables

{ Bernoulli, binomial, beta } distribution

Bernoulli distribution

R.v $x \in \{0, 1\}$, parameter μ denote the prob. of $x = 1$

$$p(x=1 \mid \underset{\text{parameter}}{\mu}) = \mu \text{ - prob.} \qquad 0 \le \mu \le 1$$

$$\mathrm{Bern}(x \mid \mu) = \mu^{x}(1-\mu)^{1-x} \qquad (x \in \{0, 1\})$$

Remark

- $p(x=0|\mu) + p(x=1|\mu) = (1-\mu) + \mu = 1$
- $\mathbb{E}[x] = 0 \cdot p(x=0|\mu) + 1 \cdot p(x=1|\mu) = \mu$
- $\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu - \mu^2 = \mu(1-\mu)$

Assume $\overset{D}{\{x_1, .. x_N\}}$ is drawn independently from some $\overset{\mathrm{Bern}(x|\mu)}{\text{Bernoulli}}$

Find a parameter μ $(= p(x=1))$ in frequentist setting

See likelihood function (of μ) ($D = \{x_1, .. x_N\}$, $x_i \in \{0, 1\}$)

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{(1-x_n)} \qquad (2.5)$$

Estimate (find) a parameter μ by maximizing (2.5)

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

$\ln p(D|\mu)$ only depends on x_n (Σx_n). Find a value for μ of

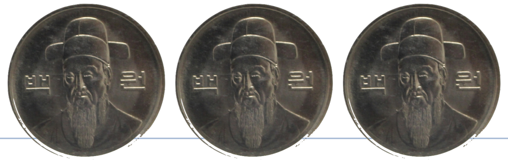
$$\frac{d}{d\mu} \ln p(D|\mu) = 0$$

$$\Rightarrow \quad \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N} \qquad \text{sample mean}$$

where m = # of $x=1$

(likelihood 를 최대화 하는 μ)

The sample mean is an example of sufficient statistic.

E.g. Flip a coin 3 times, observe 

$$\Rightarrow \quad \mu_{ML} = 1$$

Binomial distribution 이항분포

Let $x = 0$ or 1 and N be ※ trials.

R.v $m \in \{0, 1, \dots, N\}$ to be ※ of $x = 1$.

From (2.5)

$m = 0, 1, \dots$ or N

binomial distribution $\propto \mu^m (1-\mu)^{N-m}$

Ⓜ

m = ※ of $x = 1$, μ : the probability of $x = 1$

To normalize prob. dist. calculate all of possible ※ of obtaining m, $x=1$. Denote by

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

$m = 0, 1, 2, \dots N$

where $\binom{N}{m} := \frac{N!}{(N-m)!\, m!} = {}_N C_m$

We have

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

$$= \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m} = N\mu$$

(평균 횟수)

$$\mathrm{Var}[m] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \, \mathrm{Bin}(m \mid N, \mu) = N\mu(1-\mu)$$

2.1.1 The beta distribution

Conjugate distribution

$p(\theta | x)$ and $p(\theta)$ have the same distributions.

$\boxed{p(\theta)}$ is called a conjugate prior for $p(x|\theta)$

likelihood

Goal

parameter θ

x 가 어떤 분포를 따를 지 가정

⇓

posterior of θ likelihood function of θ prior of θ

Recall the $p(D|\mu)$ of Bernoulli distribution

likelihood $$p(D|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{(1-x_n)} \qquad x_n \in \{0, 1\}$$

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

To see Bayesian approach, we need to introduce $p(\mu)$.

beta distribution normalization constant

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \qquad 0 \le \mu \le 1 \qquad (2.13)$$

where $\Gamma(x) := \int_0^\infty u^{x-1} e^{-u} du$ is the gamma function (1.414)

Remark

- $\int_0^1 \mathrm{Beta}(\mu \mid a, b)\, d\mu = 1$
- $\mathbb{E}[\mu] = \dfrac{a}{a+b}$, $\mathrm{Var}[\mu] = \dfrac{ab}{(a+b)^2(a+b+1)}$ Excercise.
- a, b are called hyperparameters

The posterior dist. of μ has the form as

$$p(\mu \mid m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

where $l = N - m$. m : # of $x_n = 1$

$$\text{posterior} \propto \underbrace{p(D \mid \mu)}_{\mu^m (1-\mu)^l} \; \underbrace{p(\mu)}_{\mu^{a-1}(1-\mu)^{b-1}}$$

$$\Rightarrow \overset{\text{posterior}}{p(\mu \mid m, \ell, a, b)} = \frac{\Gamma(m+a+\ell+b)}{\Gamma(m+a)\,\Gamma(\ell+b)}\, \mu^{\overset{\text{\# of } x=1}{m}+\underline{a}-1}\,(1-\mu)^{\overset{\text{\# } x=0}{\ell}+\underline{b}-1} \qquad (2.18)$$

$$\left.\begin{array}{l} a, b : \text{ parameters of prior} \\ m, \ell : \text{ result of observation} \end{array}\right) \rightarrow \text{posterior}$$

In view of (2.18) and def of beta dist,

a, b can be interpreted as effective # observations of $x = 1$ and $x = 0$.

Sequential approach (Bayesian view point)

$x=1$ 번

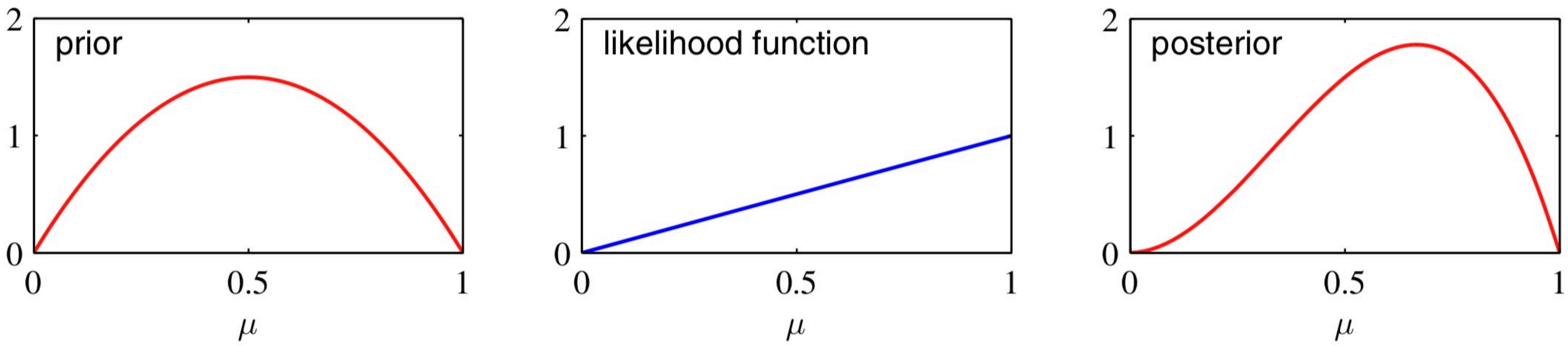


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.

Bern

시행 횟수 1

$x_1 \ldots x_N$

μ

observation

Let us predict the outcome of the next trial

$$\underline{p(x=1 \mid D)} = \int_0^1 p(x=1, \mu \mid D)\, d\mu = \int_0^1 \frac{p(x=1, \mu, D)}{p(D)}\, d\mu$$

$$= \int_0^1 \underbrace{\frac{p(x=1, \mu, D)}{p(\mu, D)} \cdot \frac{p(\mu, D)}{p(D)}}_{= \boxed{\underline{p(x=1 \mid \mu, D)}} \cdot \underline{p(\mu \mid D)}}\, d\mu$$

$A \perp B \mid C$

$\iff p(A \mid B, C)$

$= p(A \mid C)$

conditionally independent

$$\boxed{=} \int_0^1 \boxed{\underline{p(x=1 \mid \mu)}} \cdot \underline{p(\mu \mid D)}\, d\mu$$

$$= \int_0^1 \mu\, \boxed{p(\mu \mid D)}\, d\mu = \mathbb{E}_\mu[\mu \mid D]$$

posterior

$$= \frac{m+a}{m+a+l+b}$$

$m+l=N$

$$p(\underline{x=1} \mid \underline{D}) = \frac{m+a}{m+a+\ell+b} = \frac{m+a}{N+a+b} \qquad (\ell = N - m)$$

a, b are hyperparameters of prior.

m, ℓ are from the result of experiment (m = # of $x=1$)

Remark

- $m, \ell \to \infty$ (huge observations), then ML ≈ Bayesian
- $a, b \to \infty$, then variance → 0 (both prior and posterior)

2.2 Multinomial variables

Extend Bernoulli, binomial, beta distributions

Discrete variables that can take on one of K possible case

1-of-K scheme (one hot encoding)

$$\begin{array}{ccccccc} x & & \text{orange} & \text{apple} & \text{grape} & & \\ \text{orange} & & 1 & 0 & 0 & & (1,0,0)^T \\ \text{apple} & & 0 & 1 & 0 & & (0,1,0)^T \\ \text{apple} & \Rightarrow & 0 & 1 & 0 & \Rightarrow & (0,1,0)^T \\ \text{grape} & & 0 & 0 & 1 & & (0,0,1)^T \\ \text{orange} & & 1 & 0 & 0 & & (1,0,0)^T \\ & & \underline{\mu_1} & \underline{\mu_2} & \underline{\mu_3} & & \end{array}$$

k-dim. vector $\mathbb{x} = (x_1, x_2, .. x_k)^T$ as $(0,0, .. 1, 0,0..0)^T$

$$\sum_{k=1}^{K} x_k = 1$$

경우의 수 = K

$x \in \{0, 1\}$

Bernoulli ↓

Let μ_k be $p(x_k = 1)$.

The distribution of $\mathbb{x}$ (Categorical distribution)

$$p(\mathbb{x} | \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

where $\boldsymbol{\mu} := (\mu_1, .. \mu_k)^T$ with $\mu_k \geq 0$, $\sum_{k=1}^{K} \mu_k = 1$

Remark

- $\mathbf{x}$ can take K possible cases.

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

- $\mathbb{E}[\mathbf{x} | \boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x} | \boldsymbol{\mu}) \, \mathbf{x} = (\mu_1, \dots \mu_K)^T = \boldsymbol{\mu}$

Consider D of N independent observations $\mathbf{x}_1, \dots \mathbf{x}_N$.

Likelihood

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

where $m_k = \sum_n x_{nk}$ (# of observations $x_k = 1$)

D

$$\begin{array}{c|cccccc} & 1 & 2 & \cdots & k & \cdots & K \\ \mathbf{x}_1 & & & & & & \\ \mathbf{x}_2 & & & & & & \\ \vdots & & & & x_{nk} & & \\ \mathbf{x}_N & & & & & & \\ \hline \text{sum} & m_1 & m_2 & \cdots & m_k & & m_K \end{array}$$

data matrix

$$\sum_{k=1}^{K} m_k \log \mu_k$$

Maximize log likelihood $\ln p(D|\mu)$ for μ with constraint $\sum \mu_k = 1$. Using Lagrange multiplier ant maximizing

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \underbrace{\left(\sum_{k=1}^{K} \mu_k - 1 \right)}_{\text{constraint}} \qquad (2.31)$$

Set the derivative of (2.31) wrt μ_k to zero

$$\Rightarrow \qquad \mu_k = \frac{m_k}{N} \qquad\qquad \mu_{ML} = (\mu_1^{ML}, \ldots \mu_K^{ML})^T$$

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

Multinomial distribution $\mathbf{m} := (m_1, .. m_K)^T$

Joint distribution of the quantities $m_1, \dots m_K$ conditioned on $\boldsymbol{\mu}$ and on ※ of N total observations.

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

$$0 \le \mu_k \le 1$$

where $\binom{N}{m_1 m_2 \cdots m_K} := \dfrac{N!}{m_1!\, m_2! \cdots m_K!}$. $\quad \sum_{k=1}^{K} m_k = N$

K 개 categorical 변수를 갖는 N 개 자료에서 각 k-class 가 m_k 씩 가질 확률 ($\boldsymbol{\mu}$ 가 주어졌을 때)

2.2.1 Dirichlet distribution (multi - dim version of beta)

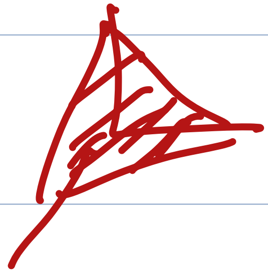
Consider the prior distributions for the parameters $\{\mu_k\}$ of multinomial distribution. (or categorical distribution)

Recall $\text{Mult}(m_1, \dots m_k \mid \mu, N) \propto \prod_{k=1}^{K} \mu_k^{m_k}$

$$\Rightarrow \qquad p(\mu \mid \alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad \underset{\mu_1}{\mu^{a-1}} \underset{\mu_2}{\boxed{(1-\mu)}^{b-1}}$$

where $0 \leq \mu_k \leq 1$, $\sum_k \mu_k = 1$. Here $\alpha := (\alpha_1, \dots \alpha_k)^T$ is the parameter

$\{\mu_k\}$ is confined to a simplex of dim $k-1$



Dirichlet distribution

$$\mathrm{Dir}(\mu \mid \alpha) := \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \underline{\prod_{k=1}^{K} \mu_k^{\alpha_k - 1}} \qquad 0 \leq \mu_k \leq 1$$

where $\Gamma(x)$ is the Gamma function, $\alpha_0 := \sum \alpha_k$

So the posterior distribution for the parameters $\{\mu_k\}$

$$\underbrace{p(\mu \mid D, \alpha)}_{\text{posterior}} \propto \underbrace{p(D \mid \mu)}_{\text{likelihood}} \overset{\text{prior}}{\underline{p(\mu \mid \alpha)}} \propto \underline{\prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}}$$

Posterior dist. again follow Dirichlet dist.

$$\prod \mu_k^{m_k} \times \prod \mu_k^{\alpha_k - 1}$$

$$\Rightarrow \quad \underline{p(\mu \mid D, \alpha)} = \mathrm{Dir}(\mu \mid \overset{\text{prior}}{\alpha} + \underset{D}{\mathbf{m}})$$

$$= \frac{\Gamma(\alpha_k + m_k)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \underline{\mu_k^{\overset{\text{Prior}}{\alpha_k} + \overset{\text{observation}}{\boxed{m_k}} - 1}}$$

where $\mathbf{m} := (m_1, \dots m_K)^T$.

As binomial dist. with beta prior, we can interprete α_k of Dirichlet prior as an effective ※ of $x_k = 1$.

2.3 The Gaussian distribution (a.k.a. normal dist.)

Single real value $x \in \mathbb{R}$

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

where μ : mean, σ^2 : variance

D - dim vector $\mathbf{x} \in \mathbb{R}^D$

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

where $\boldsymbol{\mu}$: D- dim mean vector, Σ : D×D covariance matrix

$|\Sigma|$: determinant of Σ

Remark: Why Gaussian is important?

- Fits many natural phenomena
- Maximum entropy in continuous r.v
- Central limit theorem

Fix N. iid random samples of vector $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_N$ are drawn from a population with $\boldsymbol{\mu}$, Σ.

$\Rightarrow$ R.V. $\overline{\mathbf{x}}_N := \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \approx \mathcal{N}(\boldsymbol{\mu}, \frac{1}{N}\Sigma)$

sample mean

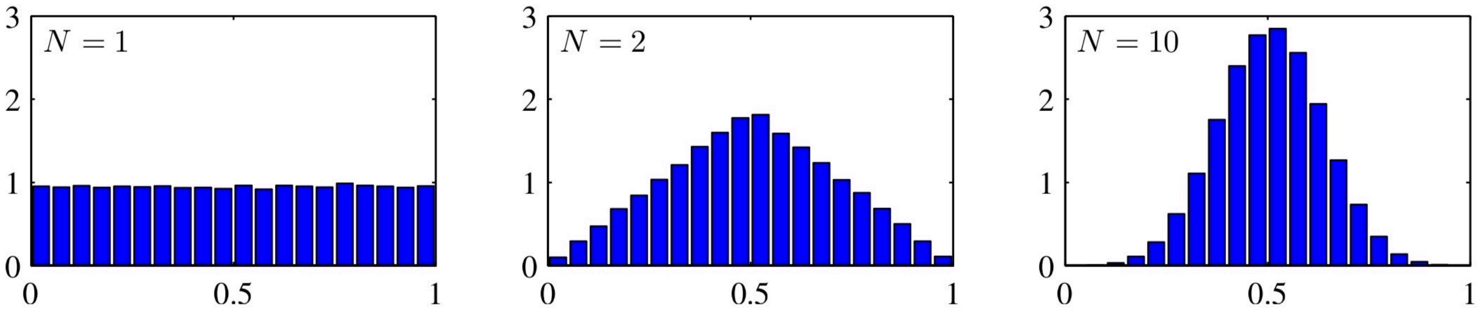


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

Let us see geometrical form of Gaussian.

$$\Delta^2 := (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \qquad \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$$

$1 \times D \qquad D \times D \qquad D \times 1$

Δ is called Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$

WLOG. assume Σ is symmetric (and real).

Δ = constant 인 $\mathbf{x}$ 의 모임 (등고선)

Consider the eigenvector equation of Σ

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

$$(\lambda_i, \mathbf{u}_i)$$

Eigenvalues $\lambda_1 \dots \lambda_D$ are real and its eigenvectors can be chosen to form an orthonormal set, so that

$$u_i^T u_j = I_{ij}$$

where I_{ij} is the i,j element of the identity matrix.

By eigenvector decomposition, Σ can be expressed as

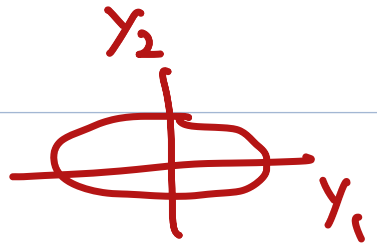
$$\underset{D \times D}{\Sigma} = \sum_{i=1}^{D} \overset{\text{value}}{\lambda_i} \, \underset{\text{vector}}{u_i u_i^T} \qquad D \times 1 \quad 1 \times D$$

and

$$\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T \qquad (\text{inverse})$$

$$\Rightarrow \qquad \Delta^2 = \underset{1\times D}{(\mathbf{x}-\boldsymbol{\mu})^T} \, \underset{D\times D}{\Sigma^{-1}} \, \underset{D\times 1}{(\mathbf{x}-\boldsymbol{\mu})} = (\mathbf{x}-\boldsymbol{\mu})^T \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \right) (\mathbf{x}-\boldsymbol{\mu})$$

Mahalanobis distance



$$= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} = C$$ 인

$\mathbf{x}$ 의 모임

where $\quad y_i := \underset{1\times D}{\mathbf{u}_i^T} \, \underset{D\times 1}{(\mathbf{x}-\boldsymbol{\mu})}$ (inner product of $\mathbf{u}_i$, $(\mathbf{x}-\boldsymbol{\mu})$)

$\mathbf{x}$

y_i : new coordinate system defined by $\mathbf{u}_i$, shifted and rotated

Let $\quad \mathbf{y} := (y_1, .., y_D)^T$. Then $\quad \mathbf{y} = U(\mathbf{x}-\boldsymbol{\mu})$ where

$$U = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_D^T \end{pmatrix}$$ whose rows are $\mathbf{u}_i^T$

(orthogonal matrix)

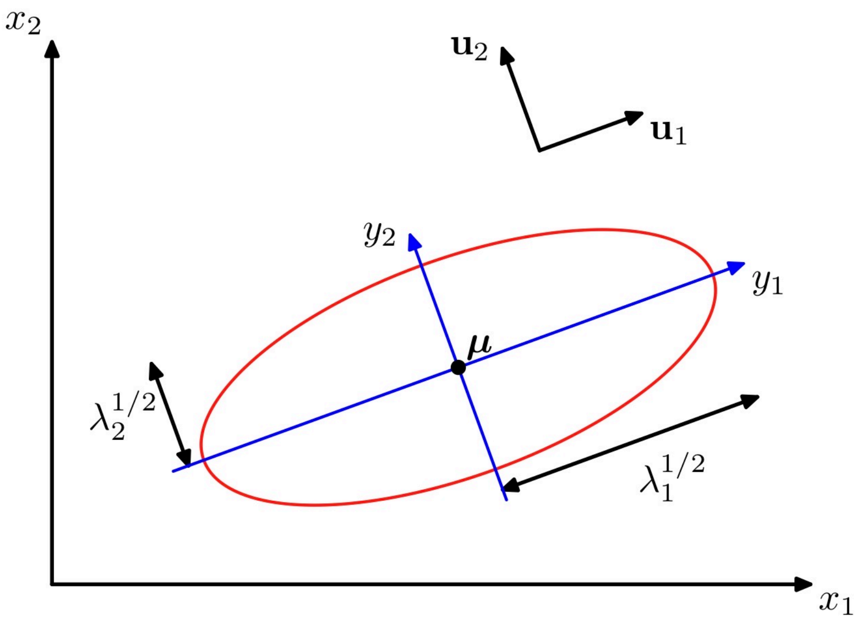
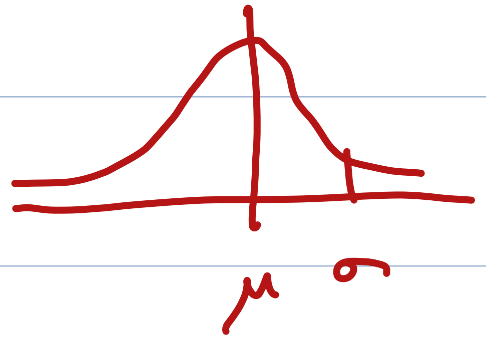


Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



Remark

- If $\lambda_i > 0$ $\forall i = 1 \ldots D$, contour surface of Δ is ellipsoid.
- center μ, axes oriented along u_i and scailing factors are given by $\lambda_i^{1/2}$

WLOG, assume all eigenvalues of Σ are strictly positive

Otherwise the distribution cannot be normalized (see ch 12)

i.e. Σ is assume to be positive definite.

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

Now consider the Gaussian dist. in y_i coordinate system.

Jacobian matrix J with

(1.27) $\underline{x = g(y)}$

$p_y(y) = p_x(g(y))\,|g'(y)|$

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}$$

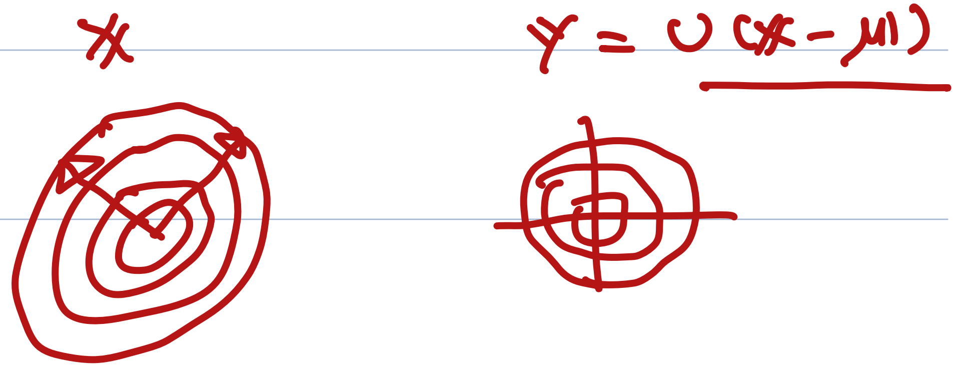
$U^T \mathbf{y} + \boldsymbol{\mu} = \mathbf{x}$

where U_{ji} are element of U^T

So $|J|^2 = |U^T|^2 = |U^T||U| = |U^T U| = |I| = 1$

and hence $|J| = \pm 1$. Also $|\Sigma|$ can be written as

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



Thus in the y_j coordinate system,

$$p(\mathbf{y}) = p(\mathbf{x})\,|\overset{\pm 1}{\det(J)}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

$$= \prod_{j=1}^{D} p_{y_j}(y_j)$$

$\mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathrm{diag}(\sqrt{\lambda_1} \ldots \sqrt{\lambda_D}))$

$\parallel$

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

In (1.49), (1.51), we found univariate Gaussian dist has

$$\mathbb{E}[x] = \mu, \quad \operatorname{var}[x] = \sigma^2$$

Now we will intepret parameters $\overset{\text{D-dim}}{\boldsymbol{\mu}}$, $\overset{D\times D}{\Sigma}$.

$$\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{2/D}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{ -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\} \mathbf{x} \, d\mathbf{x}$$

$$= \frac{1}{(2\pi)^{2/D}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{ -\frac{1}{2} \mathbf{z}^T \Sigma^{-1} \mathbf{z} \right\} (\mathbf{z}+\boldsymbol{\mu}) \, d\mathbf{z}$$

where we have changed variables using $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$.

Note that the exponent is even. So the term $\mathbf{z}$ in the factor $(\mathbf{z} + \boldsymbol{\mu})$ will vanish.

$$\Rightarrow \qquad \mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}.$$

$\mathbf{x} \in \mathbb{R}^D$

Now consider the second order moments of multivariate Gaussian. In univariate case, the second order moment is given by $\mathbb{E}[x^2]$. In multivariate Gaussian, there are D^2 second order moments given by $\mathbb{E}[x_i x_j]$.

$$\mathbb{E}[x_i x_j] \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$

$$\mathbb{E}[\underset{D\times 1}{\mathbf{x}}\,\underset{1\times D}{\mathbf{x}^T}] = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^T\,d\mathbf{x}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right\}\overset{\mathbf{z}\mathbf{z}^T}{\underbrace{(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^T}_{(*)}}\,d\mathbf{z}$$

$\mathbf{z} := \mathbf{x}-\boldsymbol{\mu}$.

$$(*)\quad (\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}^T+\boldsymbol{\mu}^T) = \underline{\mathbf{z}\mathbf{z}^T} + \overbrace{\mathbf{z}\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathbf{z}^T}^{\text{vanish by symmetry}} + \overset{\text{constant}}{\boldsymbol{\mu}\boldsymbol{\mu}^T}$$

Recall $\mathbf{y} := \underset{}{\overset{d\times d}{U}}\overset{d\times 1}{(\mathbf{x}-\boldsymbol{\mu})}$, rows of U are transpose of eigenvectors of Σ

$$\Rightarrow\quad \mathbf{z} = U^T\mathbf{y} = \underset{D\times D}{(\mathbf{u}_1 \ldots \mathbf{u}_D)}\overset{D\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_D\end{pmatrix}} = \sum_{j=1}^{D} y_j \underset{D\times 1}{\mathbf{u}_j}$$

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2} z^T \Sigma^{-1} z\right\} z z^T \, dz$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \sum_{i=1}^{D} \sum_{j=1}^{D} u_i u_j^T \int \exp\left\{-\sum_{k=1}^{D} \frac{y_k^2}{2\lambda_k}\right\} y_i y_j \, dy$$

$$= \sum_{i=1}^{D} u_i u_i^T \lambda_i = \Sigma$$

we have used $|\Sigma| = \prod_{i=1}^{D} \lambda_i$ and $\frac{1}{(2\pi)^{1/2}} \frac{1}{\lambda^{1/2}} \exp\left\{-\frac{y^2}{2\lambda}\right\} \sim \mathcal{N}(0,\lambda)$

e.g. $D=2$

$$\sum_{i} \sum_{j} \iint \exp\left(-\frac{y_1^2}{2\lambda_1}\right) \exp\left(-\frac{y_2^2}{2\lambda_2}\right) y_i y_j \, dy_1 dy_2$$

will vanish when $i \neq j$

Thus we obtain $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \Sigma$ (D×D matrix)

and covariance of $\mathbf{x}$ can be obtained by

$$\underset{D\times D}{\mathrm{cov}[\mathbf{x}]} = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T]$$

$$= \Sigma$$

Remark

$\mathcal{N}(x \mid \overset{D}{\mu}, \overset{D\times D}{\Sigma})$ sym. real

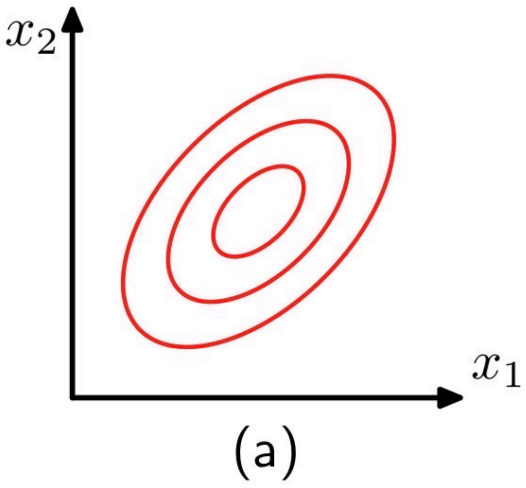
- # of parameters : $\frac{D(D+3)}{2}$ quadratic

- $\Sigma = \text{diag}(\sigma_i^2)$ or $\Sigma = \sigma^2 I$

  2D   D+1   deep dive into Gaussian

- Unimodal (single maximum)

Figure 2.8 Contours of constant probability density for a Gaussian distribution in two dimensions in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to the identity matrix, in which the contours are concentric circles.



2.3.1 Conditional Gaussian distributions

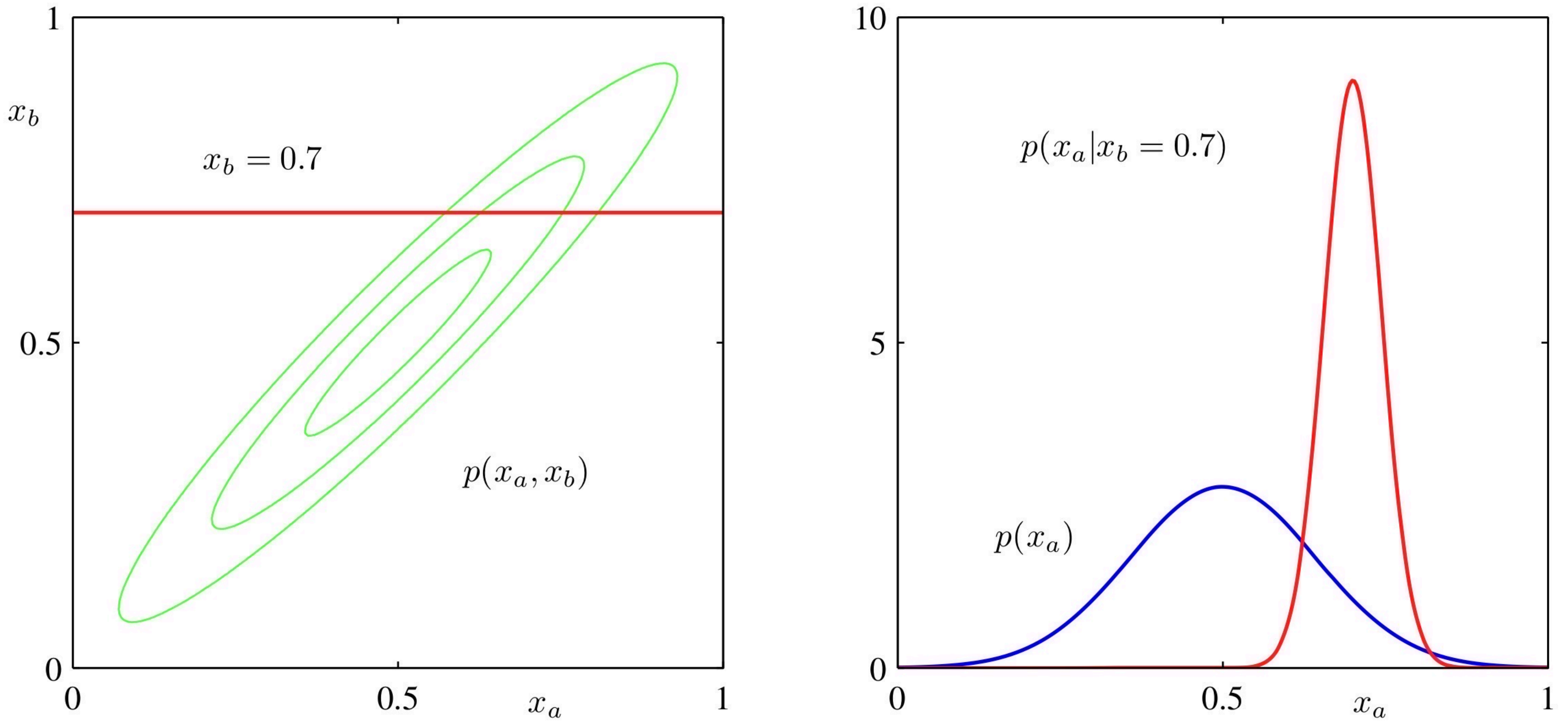


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

2.3.1 Conditional Gaussian distributions

$\mathbf{x}$: D- dimensional vector with $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$ which is partitioned into $\mathbf{x}_a$, $\mathbf{x}_b$ with $\underline{\mathbf{x}_a \in \mathbb{R}^M}$, $\underline{\mathbf{x}_b \in \mathbb{R}^{D-M}}$

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

corresponding partitions of mean vector, covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \overset{M \times M}{\Sigma_{aa}} & \overset{M \times D-M}{\Sigma_{ab}} \\ \underset{D-M \times M}{\Sigma_{ba}} & \underset{D-M \times D-M}{\Sigma_{bb}} \end{pmatrix}$$

Note that $\Sigma^T = \Sigma$ implies Σ_{aa}, Σ_{bb} are symmetric, $\Sigma_{ba} = \Sigma_{ab}^T$

Let $\Lambda := \Sigma^{-1}$. Inverse of covariance matrix, precision matrix

$$\underset{D\times D}{\Lambda} = \begin{pmatrix} \overset{M\times M}{\Lambda_{aa}} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}_{D-M\times D-M}$$

Note that Λ_{aa}, Λ_{bb} are symmetric, $\Lambda_{ba} = \Lambda_{ab}^{T}$.

$\Lambda_{aa} \neq \Sigma_{aa}^{-1}$ excercise

Find conditional distribution $p(\mathbf{x}_a \mid \mathbf{x}_b)$. Fix $\mathbf{x}_b$.

Consider the quadratic form in the exponent.

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \qquad \underline{1\times D} \quad \underline{D\times D} \quad \underline{D\times 1} \qquad (\mathbf{x}_b \text{ fixed})$$

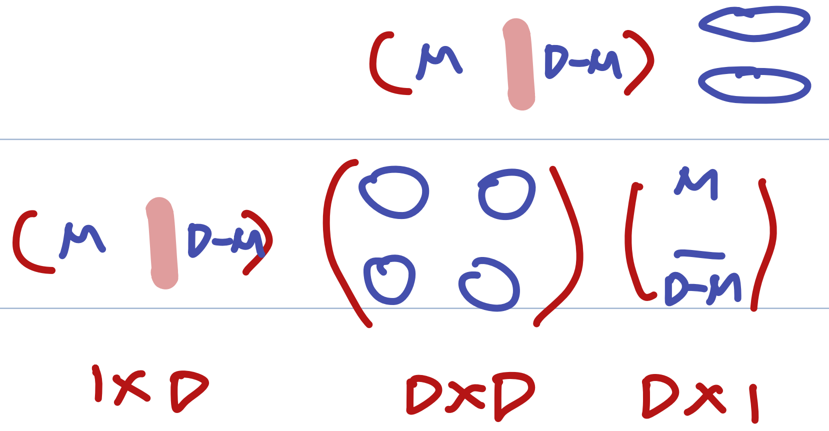
$\mathbf{x}_a, \mathbf{x}_b$

$$= -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\Lambda_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T\Lambda_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b) \qquad (2.70)$$

First, $p(\mathbf{x}_a|\mathbf{x}_b)$ will be M-dim Gaussian, because density function is a quadratic form of exponent.

Now we are going to find its mean vector and covariance by "completeting the square".



E.g. in case $p(x) \propto \exp(ax^2+bx+c)$ → x 는 gaussian

① ⇒ $p(x) \propto \exp\left\{a\left(x^2+\frac{b}{a}x+\frac{b^2}{4a^2}\right)-\frac{b^2}{4a}+c\right\}$

$$\propto \exp\left\{a\left(x+\frac{b}{2a}\right)^2\right\}$$

② $p(x) = \mathcal{N}(x \mid \mu, \sigma^2)$ $\mu = -\frac{b}{2a}$, $\sigma^2 = -\frac{1}{2a}$

Since $\mathcal{N}(x \mid \mu, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\right\}$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x\right\}$$

⇒ $a = -\frac{1}{2\sigma^2}$, $b = \frac{\mu}{\sigma^2}$ $\left(\mu = -\frac{b}{2a},\ \sigma^2 = -\frac{1}{2a}\right)$

Likewise, the exponent in D-dim Gaussian can be written

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \underbrace{\boxed{-\frac{1}{2}}\mathbf{x}^T \underline{\Sigma^{-1}}\mathbf{x}} + \mathbf{x}^T \underline{\Sigma^{-1}\boldsymbol{\mu}} + \text{constant}$$

In view of (2.70), (second order in $\mathbf{x}_a$)

$$-\frac{1}{2}\mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a$$

So we obtain covariance of $p(\mathbf{x}_a | \mathbf{x}_b)$ is given by

$$\Sigma_{a|b} := \Lambda_{aa}^{-1} \quad (\neq \Sigma_{aa})$$

$\boldsymbol{\mu}_{a|b}$?

Now consider all of the terms in (2.70) that are linear in $\mathbf{x}_a$

$$\mathbf{x}_a^T \{ \Lambda_{aa} \boldsymbol{\mu}_a - \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \}$$

where we have used $\Lambda_{ba}^T = \Lambda_{ab}$.

Since $\Sigma_{a|b}^{-1} \boldsymbol{\mu}_{a|b} = \Lambda_{aa} \boldsymbol{\mu}_a - \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b)$,

$$\boldsymbol{\mu}_{a|b} = \overset{\Lambda_{aa}^{-1}}{\overset{\shortparallel}{\Sigma_{a|b}}} \{ \Lambda_{aa} \boldsymbol{\mu}_a - \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \}$$

$$= \boldsymbol{\mu}_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\therefore \quad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b), \qquad \Sigma_{a|b} = \Lambda_{aa}^{-1}$$

Let us find Λ_{aa} and Λ_{ab}.

Recall $\Lambda = \Sigma^{-1}$, $\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$

Use the following identification for the inverse of a partitioned matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \qquad (2.76)$$

where $M := (A - BD^{-1}C)^{-1}$

So we have

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$

$$\Lambda_{ba} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

and hence

$$\left\{\begin{array}{l} \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \qquad x_b \text{ fix} \\ \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \end{array}\right.$$

Remark

- $\mu_{a|b}$ is a linear function of x_b
- $\Sigma_{a|b}$ is independent of x_b

2.3.2 Marginal Gaussian distributions

Consider the following marginal distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)$

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, d\mathbf{x}_b \qquad \mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \begin{matrix} M \\ D-M \end{matrix}$$

($\mathbf{x}_b$ 로 주변화하고 남은 $\mathbf{x}_a$는 어떤 분포인지)

Strategy : focus on quadratic form of exponent and identify the mean vector and covariance matrix of $p(\mathbf{x}_a)$

Recall (2.70) $\quad \Sigma^{-1} = \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$

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

$$= -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T \Lambda_{aa} (\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^T \Lambda_{ab} (\mathbf{x}_b-\boldsymbol{\mu}_b)$$

$$-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T \Lambda_{ba} (\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^T \Lambda_{bb} (\mathbf{x}_b-\boldsymbol{\mu}_b)$$

In order to integrate out $\mathbf{x}_b$, pick out those terms involving $\mathbf{x}_b$

$$-\frac{1}{2}\mathbf{x}_b^T \Lambda_{bb}\mathbf{x}_b + \mathbf{x}_b^T \mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})^T \Lambda_{bb}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})$$

(2.84)

$$+ \underbrace{\frac{1}{2}\mathbf{m}^T \Lambda_{bb}^{-1}\mathbf{m}}_{\text{indep. of } \mathbf{x}_b} \qquad \text{(square expression)}$$

where $\mathbf{m} := \Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)$

For $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, d\mathbf{x}_b$,

(2.86) $$\int \exp\left\{-\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})^T \Lambda_{bb}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})\right\} d\mathbf{x}_b$$

which is an inverse of the normalization coefficient.

As seen before, this coefficient is independent of mean.

Combining the last term ($\frac{1}{2}\mathbf{m}^T \Lambda_{bb}^{-1} \mathbf{m}$) in (2.84) with remaining terms in (2.70) depending on $\mathbf{x}_a$, we obtain

$$-\frac{1}{2}\mathbf{m}^T \Lambda_{bb}^{-1} \mathbf{m} - \frac{1}{2}\mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + \mathbf{x}_a^T (\Lambda_{aa}\boldsymbol{\mu}_a + \Lambda_{ab}\boldsymbol{\mu}_b) + \text{constant}$$

$$= -\frac{1}{2}[\Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)]^T \Lambda_{bb}^{-1} [\Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)]$$

$$+ \mathbf{x}_a^T (\Lambda_{aa}\boldsymbol{\mu}_a + \Lambda_{ab}\boldsymbol{\mu}_b) + \text{constant}$$

$$= -\frac{1}{2}\mathbf{x}_a^T (\underline{\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}})\mathbf{x}_a + \mathbf{x}_a^T (\underline{\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}})\boldsymbol{\mu}_a$$

$$+ \text{constant}$$

Recall the exponent in D-dim Gaussian can be written

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

Denote the covariance of $p(\mathbf{x}_a)$ by Σ_a and Σ_a is given by

$$\Sigma_a = \underbrace{(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})}_{(\Sigma_a)^{-1}}{}^{-1}$$

Similarly, mean vector is given by

$$\Sigma_a\underbrace{(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})}_{\Sigma_a^{-1}}\boldsymbol{\mu}_a = \boldsymbol{\mu}_a$$

To simplify $\Sigma_a = (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba})^{-1}$,

recall

$$\begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

and use (2.76) expression of the inverse of a partitioned matrix

$$\Sigma_a = (\Lambda_{aa} - \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba})^{-1} = \underline{\Sigma_{aa}}$$

$$\Sigma = \begin{pmatrix} \boxed{\Sigma_{aa}} & \\ & \end{pmatrix}$$

Thus we have $\mathbb{E}[\mathbf{x}_a] = \underline{\mu_a}$, $\operatorname{cov}[\mathbf{x}_a] = \underline{\Sigma_{aa}}$

where $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, d\mathbf{x}_b$

$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$ with $\Lambda := \Sigma^{-1}$ D-dim $\mathbf{x}$

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

Conditioned distribution

$$p(\mathbf{x}_a \mid \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \Lambda_{aa}^{-1})$$

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

Marginal distribution

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_a, \Sigma_{aa})$$

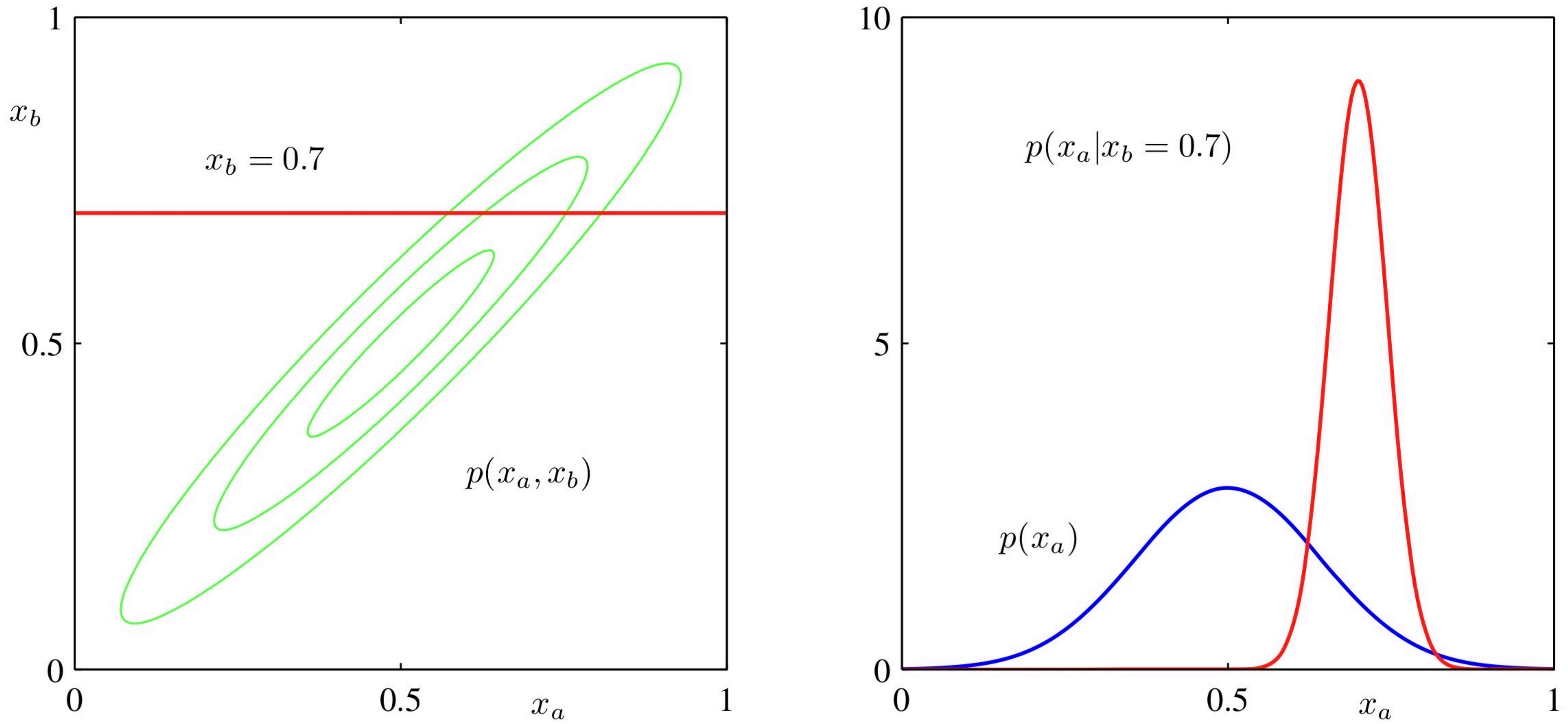


Figure 2.9 The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

2.3.3 Bayes' theorem for Gaussian variables

Linear Gaussian model example

Gaussian marginal dist. $p(\mathbf{x})$, Gaussian conditional dist $p(\mathbf{y}|\mathbf{x})$

$p(\mathbf{y}|\mathbf{x})$ has a mean as a linear function of $\mathbf{x}$ and a covariance which is independent of $\mathbf{x}$.

i.e.

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

$\mathbf{x} \in M$-dim

$\mathbf{y} \in D$-dim

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

where $\boldsymbol{\mu}$, A and $\mathbf{b}$ are parameters governing the means, and Λ and L are precision matrices.

We will find $\underset{\text{marginal}}{p(\mathbf{y})}$ and $\underset{\text{conditional}}{p(\mathbf{x}|\mathbf{y})}$.

known $p(\mathbf{x})$

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

Let $\mathbf{z} := \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ and us consider the joint prob. dist

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

$$\ln p(\mathbf{z}) = \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \qquad (2.102)$$

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

$$-\frac{1}{2}(\mathbf{y}-A\mathbf{x}-\mathbf{b})^T L (\mathbf{y}-A\mathbf{x}-\mathbf{b}) + \overset{\text{indep. of } \mathbf{x},\, \mathbf{y}.}{\text{const}}$$

This is a quadratic function of the $\overset{\mathbf{x},\,\mathbf{y}}{\text{component}}$ of $\mathbf{z}$,

hence $p(\mathbf{z})$ is a Gaussian

Consider the second term in (2.102)

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

$$= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\Lambda + A^T L A & -A^T L\\ -LA & L\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T R\, \mathbf{z}$$

$\mathbf{z}$ has precision (inverse of covariance) matrix given by

$$\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \qquad \begin{pmatrix}\Lambda + A^T L A & -A^T L\\ -LA & L\end{pmatrix}$$

$$\Rightarrow \quad \mathrm{cov}[\mathbf{z}] = R^{-1} = \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T\\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T\end{pmatrix} \qquad (2.105)$$

Consider the linear term in (2.102)

$$\mathbf{x}^T \Lambda \boldsymbol{\mu} - \mathbf{x}^T A^T L \mathbf{b} + \mathbf{y}^T L \mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \Lambda \boldsymbol{\mu} - A^T L \mathbf{b} \\ L \mathbf{b} \end{pmatrix}$$

$$\Rightarrow \qquad \mathbb{E}[\mathbf{z}] = R^{-1} \begin{pmatrix} \Lambda \boldsymbol{\mu} - A^T L \mathbf{b} \\ L \mathbf{b} \end{pmatrix}$$

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \qquad = \left(\begin{array}{c} \boldsymbol{\mu} \\ A\boldsymbol{\mu} + \mathbf{b} \end{array} \right) \begin{array}{c} \mathbf{x} \\ \mathbf{y} \end{array} \qquad (2.108)$$

$$A\mathbf{x} + b$$

Using section 2.3.2 and $p(\mathbf{y}) = \int p(\mathbf{z})\, d\mathbf{x}$,

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

$$\mathrm{cov}[\mathbf{y}] = L^{-1} + A \Lambda^{-1} A^T$$

Now we can find an expression for $p(\mathbf{x} | \mathbf{y})$.

$$\mathbb{E}[\mathbf{x} | \mathbf{y}] = (\Lambda + A^T L A)^{-1} \{ A^T L (\mathbf{y} - \mathbf{b}) + \Lambda \boldsymbol{\mu} \}$$

$$\mathrm{cov}[\mathbf{x} | \mathbf{y}] = (\Lambda + A^T L A)^{-1}$$

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) & (2.113) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) & (2.114) \end{aligned}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) & (2.115) \\ p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) & (2.116) \end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \qquad (2.117)$$

2.3.4 Maximum likelihood for the Gaussian

Data set $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_N)^T$, $\{\mathbf{x}_n\}$ iid samples of D-dimensional Gaussian. The log likelihood function is given by

$\ln p(\mathbf{X} | \boldsymbol{\mu}, \Sigma)$ $\mathbf{X}$ $N \times D$ matrix

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

Note that likelihood function depends only on the following two quantities

$$\sum_{n=1}^{N} \mathbf{x}_n \quad , \quad \sum_{n=1}^{N} \underbrace{\mathbf{x}_n \mathbf{x}_n^T}_{}$$

$D \times D$

These are known as sufficient statistics for Gaussian

$$\nabla_{\mu} \ln p(X \mid \mu, \Sigma) = \underline{\sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu)}$$ D - dim vector

Set this gradient to zero vector, we obtain

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

solution of maximum likelihood estimator

MLE

(sample mean)

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

(sample covariance)

Remark

- Σ_{ML} involves μ_{ML}
- μ_{ML} is independent of Σ_{ML}

Evaluate the expectations of thes solutions under the true distribution. Then we obtain

$$\mathbb{E}[\mu_{ML}] = \mu \qquad \text{unbiased estimate}$$

$$\mathbb{E}[\Sigma_{ML}] = \frac{N-1}{N}\Sigma \qquad \text{biased}$$

2.3.5 Sequential estimation

Sequential estimation for maximum likelihood

This method allows data points to be proceed one at time and then discarded and are important for on-line applications

Consider

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

which we will denote by $\mu_{ML}^{(N)}$ based on N observations

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

$$= \frac{1}{N} x_N + \frac{1}{N}\sum_{n=1}^{N-1} x_n$$

$$= \frac{1}{N} x_n + \frac{N-1}{N} \mu_{ML}^{(N-1)}$$

$$= \mu_{ML}^{(N-1)} + \frac{1}{N}\underbrace{\left(x_N - \mu_{ML}^{(N-1)} \right)}_{\text{error signal}}$$

General formulation of sequential learning (Robbins - Monro)

Two r.v. z, θ governed by a joint distribution $p(z,\theta)$

Define deterministic function $f(\theta)$ by

$$f(\theta) := \mathbb{E}_z[z|\theta] = \int z\, p(z|\theta)\, dz$$

coditional expectation

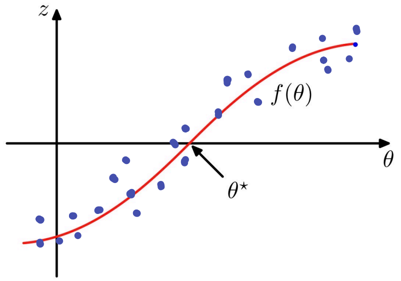
e.g. $\mathbb{E}[t|x]$

which is a function of θ (called regression function)

Find the root θ^* at which $f(\theta^*) = 0$

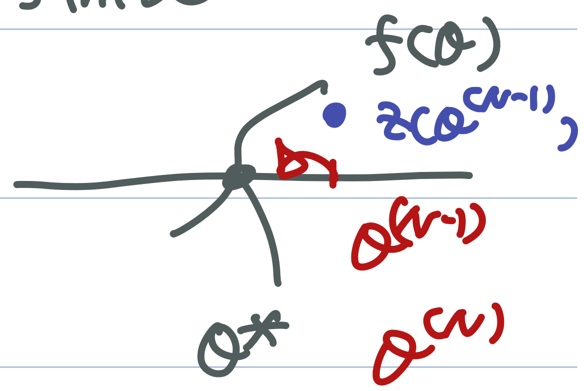
Suppose we can observe values of z one at a time

Figure 2.10 A schematic illustration of two correlated random variables z and θ, together with the regression function $f(\theta)$ given by the conditional expectation $\mathbb{E}[z|\theta]$. The Robbins-Monro algorithm provides a general sequential procedure for finding the root θ^* of such functions.



Assume the conditional variance of z is finite

$$\mathbb{E}_z[(z-f)\,|\,\theta] < \infty$$



θ^* solution

and wlog $f(\theta) > 0$ for $\theta > \theta^*$ and $f(\theta) < 0$ for $\theta < \theta^*$

A sequence of successive estimates of the root θ^* given by

$$\theta^{(N)} := \theta^{(N-1)} - a_{N-1}\, z(\theta^{(N-1)}) \qquad (2.129)$$

where $z(\theta^{(N)})$ is an observed value of z when $\theta = \theta^{(N)}$

$\{a_N\}$ represents a seq of positive numbers satisfying

$$\lim_{N\to\infty} a_N = 0$$

$$\sum_{N=1}^{\infty} a_N = \infty \qquad \text{e.g. } \frac{1}{N}$$

$$\sum_{N=1}^{\infty} a_N^2 < \infty$$

of $f(\theta)$

By [Robbins - Monro], (2.129) converges to the root with probability one.

Remark

- Third condition ensures that the accumulated noise has finite variance and hence does not spoil convergence.

General Maximum likelihood problem

$$f(\theta) = \int z \, p(z|\theta) \, dz$$

$$\|$$

$$\mathbb{E}_x\left[\frac{\partial}{\partial\theta} \ln p(x|\theta)\right]$$

By definition of θ_{ML}, θ_{ML} satisfies

$$\frac{\partial}{\partial\theta}\left\{ -\frac{1}{N}\sum_{n=1}^{N} \ln p(x_n|\theta) \right\}\Bigg|_{\theta_{ML}} = 0$$

Taking $N \to \infty$ and exchanging derivative and summation,

$$-\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial\theta} \ln p(x_n|\theta) = \mathbb{E}_x\left[-\frac{\partial}{\partial\theta} \ln p(x|\theta)\right]$$

$$\mathbb{E}_x[f(x)] \approx \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$ observations x_n

I.e. find the root of a regression function

Apply Robbins - Monro procedure

$$\theta^{(N)} := \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} [-\ln p(x_N | \theta^{(N-1)})] \quad (2.135)$$

Specific example: sequential estimation of the mean of Gaussian distribution

In this case $\theta^{(N)}$ is the $\mu_{ML}^{(N)}$ mean of the Gaussian and z is given by

$$z = \frac{\partial}{\partial \mu_{ML}} \ln p(x | \mu_{ML}, \sigma^2) = -\frac{1}{\sigma^2}(x - \mu_{ML}) \quad (2.136)$$

Substituting (2.136) into (2.135) with $a_N = \sigma^2 / N-1$

then we obtain (2.126)

2.3.6 Bayesian inference for the Gaussian

MLE method gave point estimates for μ, Σ (section 2.3.4)

Now develop a Bayesian treatment ①

Single Gaussian random variable x. Suppose σ^2 is known.

Aim to inference μ given N observations $\mathcal{X} = \{x_1, \dots x_N\}$

The likelihood function is given by

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

Note that this function is the form of the exponential of a quadratic form of μ.

We will choose a prior $p(\mu)$ given by Gaussian because the product of two exponentials of quadratic function of μ will also be Gaussian

Take prior prob. $p(\mu)$ to be

μ_0, σ_0^2 hyperparameters

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

Posterior

$$p(\mu \mid \mathbf{x}) \propto p(\mathbf{x} \mid \mu) \cdot \overset{\mathcal{N}(\mu \mid \mu_0, \sigma_0^2)}{\overset{\shortparallel}{p(\mu)}}$$

Exercise 2.38. we obtain

posterior parameters

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

$\frac{1}{N} \sum x_n$

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

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

Remark

- μ_N is a compromise between μ_0 and μ_{ML}
- Effect of change in value N

- Precision is additive, if $N \to \infty$, $\sigma_N^2 \to 0$
- When N is finite, if $\sigma_0^2 \to \infty$, then the poterior mean reduces to μ_{ML} and variance σ_N^2 becomes $\frac{\sigma^2}{N}$

Sequential inference in Bayesian paradigm

$$\underbrace{p(\mu \mid \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\left[p(\mu) \prod_{n=1}^{N-1} p(x_n \mid \mu) \right]}_{\propto \text{ posterior distribution after observing } N-1 \text{ data}} p(x_N \mid \mu)$$

$\mathbf{X} = \{x_1 \dots x_N\}$

② Now we wish to infer the variance and assume mean is known.

Let the precision $\lambda := 1/\sigma^2$. The likelihood function of λ

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

$\lambda > 0$

i.e. the form of $\lambda^{N/2} \cdot \exp(-\frac{\lambda}{2})$

The corresponds to gamma distribution

$a, b > 0$

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^{a} \lambda^{a-1} \exp(-b\lambda),$$

$\underline{\lambda > 0}$

Here $\Gamma(a)$ is a gamma function $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$

Remark

- If $a>0$, gamma distribution has finite integral.
- If $a \geq 1$, the distribution itself is finite.
- $\mathbb{E}[\lambda] = \frac{a}{b}$ $\quad$ $\mathrm{var}[\lambda] = \frac{a}{b^2}$

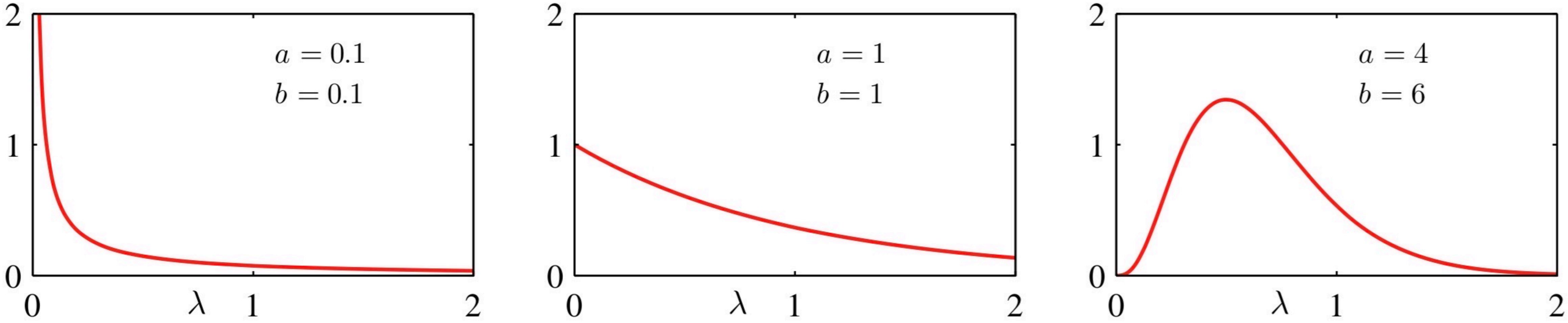


Figure 2.13 Plot of the gamma distribution $\mathrm{Gam}(\lambda|a, b)$ defined by (2.146) for various values of the parameters a and b.

Consiter the prior dist. $\mathrm{Gam}(\lambda | a_0, b_0)$. ($a_0, b_0$: hyperparameter)

The posterior dist. of λ is as below

$$p(\lambda|\mathbf{x}) \propto \underbrace{\lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}}_{\text{likelihood function of } \lambda} \cdot \underbrace{\mathrm{Gam}(\lambda|a_0,b_0)}_{\text{prior of } \lambda}$$

$$\propto \lambda^{a_0-1}\lambda^{N/2}\exp\left\{-b_0\lambda-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \qquad (2.149)$$

$\Rightarrow$ $p(\lambda|\mathbf{x}) = \mathrm{Gam}(\lambda|a_N, b_N)$ where

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

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

Remark

- Effect of observing N data points
  - increases the value of a by $\frac{N}{2}$
  - " b by $\frac{N}{2}\sigma_{ML}^2$
- We can interprete the parameter a_0 in terms of $2a_0$ 'effective' prior observations.
- $\mathbb{E}[\lambda \mid \mathbf{x}] = \dfrac{a_N}{b_N} = \dfrac{2a_0 + N}{2b_0 + N\sigma_{ML}^2}$

$\lambda = \dfrac{1}{\sigma^2}$

$\mathrm{var}[\lambda \mid \mathbf{x}] = \dfrac{a_N}{b_N^2}$ $\quad\mathbb{E}[\lambda] \longrightarrow \dfrac{1}{\sigma_{ML}^2}$

Now suppose that the both μ and λ are unknown

Consider the dependence of the likelihood function on μ and λ

$$p(\mathbf{x} \mid \mu, \lambda) = \prod_{n=1}^{N} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n - \mu)^2\right\}$$

$$\lambda, \mu \qquad \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda \mu^2}{2}\right)\right]^N \exp\left[\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2\right]$$

Thus the prior distribution should take the form

$$p(\mu, \lambda) \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta} \exp\{c\lambda\mu - d\lambda\}$$

(2.153)

$$= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}}_{\mu,\,\lambda} \underbrace{\lambda^{\beta/2} \exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}}_{\lambda}$$

where c, d and β are constants. Use $p(\mu, \lambda) = p(\mu | \lambda)\, p(\lambda)$.

$p(\mu|\lambda)$: a Gaussian whose precision is a linear function of λ

$p(\lambda)$: a gamma distribution. So we take a prior

$$p(\mu, \lambda) = \mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1})\, \mathrm{Gam}(\lambda \mid a, b) \qquad (2.154)$$

where $\mu_0 := c/\beta$, $a := (1+\beta)/2$, $b := d - c^2/2\beta$

(2.154) is called normal gamma or Gaussian gamma.

Note that it is not the simply the product of an independent Gaussian prior and gamma prior.

Multivariate Gaussian $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \Lambda^{-1})$ for D-dim $\mathbf{x}$

First, when precision matrix Λ is known, the conjugate prior distribution is again a Gaussian.

Second, for known mean and unknown precision matrix $\overset{D\times D}{\Lambda}$, the conjugate prior distribution is the Wishart distribution given by

$$\mathcal{W}(\Lambda | \overset{D\times D}{\mathbf{W}}, \nu) = B|\Lambda|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}\overset{\text{trace of matrix}}{\mathrm{Tr}}(\mathbf{W}^{-1}\Lambda)\right)$$

where ν is called the number of degrees of freedom

W is a $D \times D$ scale matrix. The nomalization constant B is given by

$$B(W, \nu) = |W|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}.$$

If both the mean ant precision are unknown, the conjugate prior is given by

$$p(\mu, \Lambda \mid \mu_0, \beta, W, \nu) = \mathcal{N}(\mu \mid \mu_0, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda \mid W, \nu)$$

which is known as the normal-Wishart or Gaussian-Wishart

2.3.7 Student's t - distribution

Conjugate prior for the precision of a Gaussian is given by a gamma distribution.

Consider univariate Gaussian $\mathcal{N}(x \mid \mu, \tau^{-1})$ with Gamma prior $\mathrm{Gam}(\tau \mid a, b)$. Integrate out the precision

$$p(x \mid \mu, a, b) = \int_0^\infty \mathcal{N}(x \mid \mu, \tau^{-1}) \, \mathrm{Gam}(\tau \mid a, b) \, d\tau \qquad \tau > 0$$

$$= \int_0^\infty \frac{b^a e^{(-b\tau)} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi} \right)^{1/2} \exp\left\{ -\frac{\tau}{2} (x-\mu)^2 \right\} d\tau$$

$$= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi} \right)^{1/2} \left[b + \frac{(x-\mu)^2}{2} \right]^{-a-1/2} \Gamma(a + 1/2)$$

where we have made the change of variable $z = \tau[b + (x-\mu)^2/2]$

Define new parameters $\nu = 2a$ and $\lambda = a/b$.

$$St(x \mid \mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2 - 1/2}$$

known as Student's t-distribution. λ is called precision and ν is called the degree of freedom.

When $\nu = 1$, t-distribution reduces to the Cauchy dist.

While in the limit $\nu \to \infty$, t-distribution becomes Gaussian $N(x \mid \mu, \lambda^{-1})$

Remark

- t-dist. can be interpreted infinite mixture of Gaussian

- Longer tail, robustness property

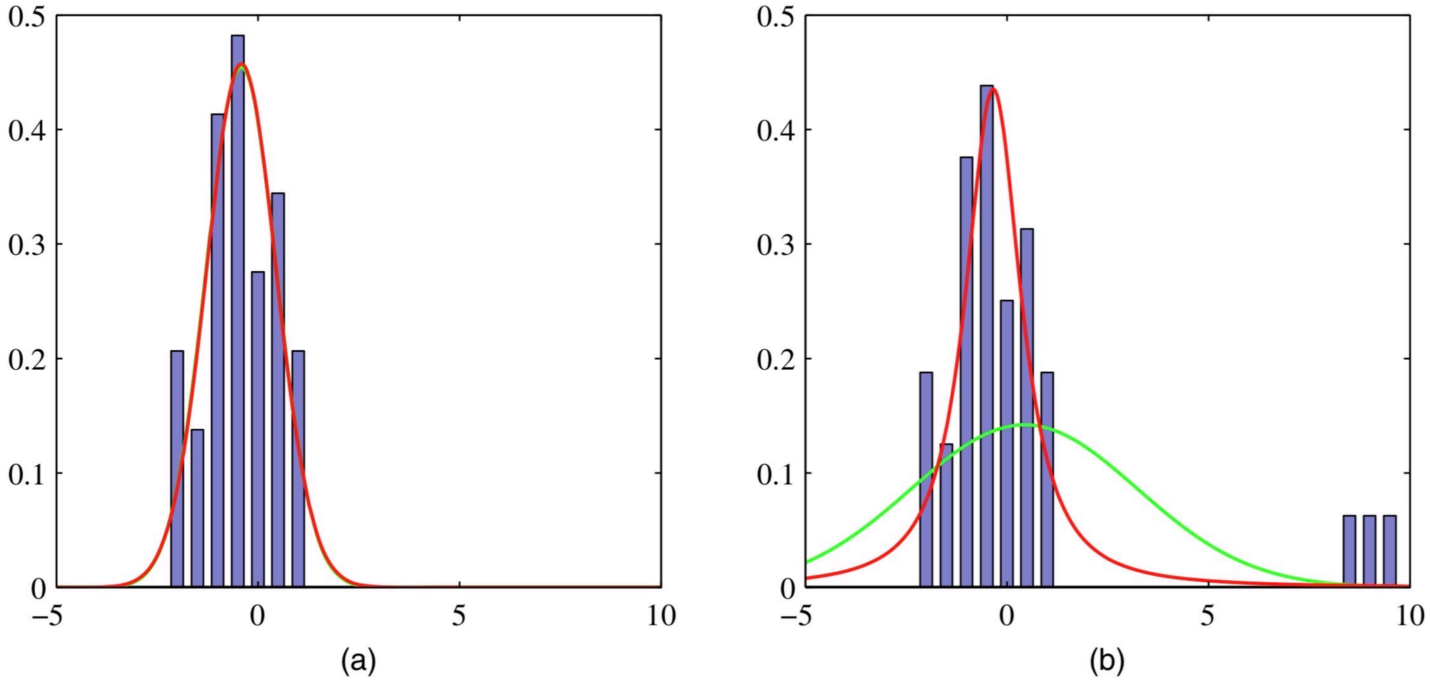


Figure 2.16 Illustration of the robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden by the red curve). Because the t-distribution contains the Gaussian as a special case it gives almost the same solution as the Gaussian. (b) The same data set but with three additional outlying data points showing how the Gaussian (green curve) is strongly distorted by the outliers, whereas the t-distribution (red curve) is relatively unaffected.

Multivariate Student's t - distribution

$$\mathrm{St}(\mathbf{x} \mid \boldsymbol{\mu}, \Lambda, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, (\eta \Lambda)^{-1}) \, \mathrm{Gam}(\eta \mid \nu/2, \nu/2) \, d\eta$$

$$= \frac{\Gamma(D/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{1/2}}{(\pi \nu)^{D/2}} \left[1 + \frac{\Delta^2}{\nu} \right]^{-D/2 - \nu/2}$$

where $\Delta := (\mathbf{x} - \boldsymbol{\mu})^T \Lambda (\mathbf{x} - \boldsymbol{\mu})$

Remark

- $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ if $\nu > 1$
- $\mathrm{cov}[\mathbf{x}] = \frac{\nu}{(\nu - 2)} \Lambda^{-1}$ if $\nu > 2$
- $\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu}$

2.3.8 Periodic variables

Consider an angular (polar) coordinate $0 \leq \theta < 2\pi$ and the problem of evaluating the mean of observations $D = \{\theta_1, \ldots \theta_N\}$.

Simple average $(\theta_1 + \cdots \theta_N)/N$ is strongly coordinate depentent.

Set angular observations as points on unit circle.

Let x_i be two-dim vector with $x_i = (\cos\theta_i, \sin\theta_i)$

Average the vectors $\{\mathbf{x}_n\}$ instead to give

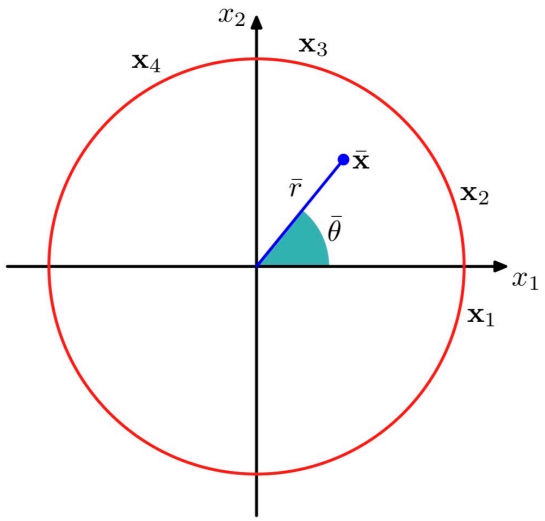


Figure 2.17 Illustration of the representation of values θ_n of a periodic variable as two-dimensional vectors $\mathbf{x}_n$ living on the unit circle. Also shown is the average $\bar{\mathbf{x}}$ of those vectors.

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$

$$= \bar{r} \left(\cos\bar{\theta}, \sin\bar{\theta} \right)$$

i.e. $\bar{r}\cos\bar{\theta} = \frac{1}{N} \sum_{n=1}^{N} \cos\theta_n$, $\bar{r}\sin\bar{\theta} = \frac{1}{N} \sum_{n=1}^{N} \sin\theta_n$

Thus we can solve for $\bar{\theta}$ to give

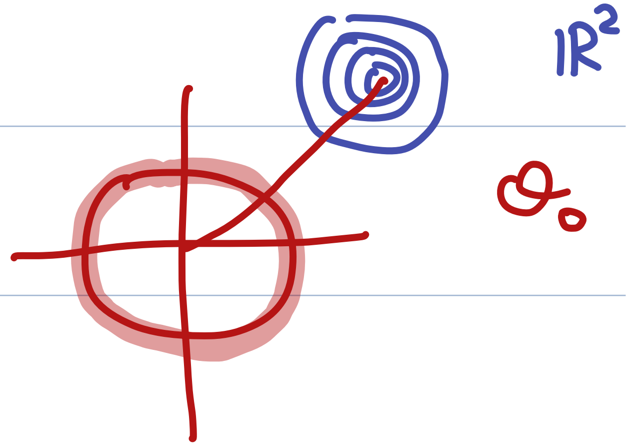
$$\bar{\theta} = \tan^{-1}\left\{ \frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} \right\}$$

Consider $p(\theta)$ that have period 2π and must satisfies

$$p(\theta) \geq 0$$

$$\int_0^{2\pi} p(\theta)\, d\theta = 1$$

$$p(\theta + 2\pi) = p(\theta)$$



We can easily obtain a Gaussian - like distribution.

Consider a Gaussian over $\mathbf{x} = (x_1, x_2)$ having mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and covariance matrix $\Sigma = \sigma^2 I$ so that

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2} \right\} \qquad (2.173)$$

Map $\mathbf{x} = (x_1, x_2)$ and $\boldsymbol{\mu}$ into polar coordinates

$$\begin{cases} x_1 = r\cos\theta, & x_2 = r\sin\theta \\ \mu_1 = r_0\cos\theta_0, & \mu_2 = r_0\sin\theta_0 \quad \text{fixed } r_0, \theta_0 \end{cases}$$

Substitute these transformation into (2.173) with $r=1$ condition

The exponent in (2.173)

$$-\frac{1}{2\sigma^2}\left\{(r\cos\theta - r_0\cos\theta_0)^2 + (r\sin\theta - r_0\sin\theta_0)^2\right\}$$

$(r=1)$

$$= -\frac{1}{2\sigma^2}\left\{1 + r_0^2 - 2r_0\cos\theta\cos\theta_0 - 2r_0\sin\theta\sin\theta_0\right\}$$

$$= \frac{r_0}{\sigma^2}\cos(\theta - \theta_0) + \text{const}$$

Define $m = r_0/\sigma^2$. Then we obtain the expression for the distribution of $p(\theta)$ along unit circle

$$p(\theta \mid \theta_0, m) = \frac{1}{2\pi}\frac{1}{I_0(m)} \exp\{ m \cos(\theta - \theta_0) \} \qquad 0 \leq \theta \leq 2\pi$$

//

which called von Mises distribution. Here θ_0 represents the mean and $m = r_0/\sigma^2$ is called concentration parameter.

$I_0(m)$: zeroth - order Bessel function of the first kind

$$I_0(m) := \frac{1}{2\pi}\int_0^{2\pi} \exp\{ m \cos\theta \} \, d\theta$$

Now consider the maximum likelihood for θ_0 and m

Observations $D = \{\theta_1, \dots \theta_N\}$ is given

$$\ln p(D \mid \theta_0, m) = \prod_{n=1}^{N} p(\theta_n \mid \theta_0, m) \qquad (2.181)$$

$$= -N \ln(2\pi) - N \ln I_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0)$$

Set the derivative w.r.t θ_0 equal to zero gives

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0$$

Thus, we obtain $\theta_0^{ML} = \tan^{-1}\left\{\frac{\sum \sin\theta_n}{\sum \cos\theta_n}\right\}$

Similarly maximizing (2.181) w.r.t m. Set the derivative of (2.181) w.r.t m, then we have

$$A(m) = \frac{1}{N}\sum_{n=1}^{N} \cos(\theta_n - \theta_0^{ML}) \qquad (2.185)$$

where we used $I_0'(m) = I_1(m)$ and have defined

$$A(m) := \frac{I_1(m)}{I_0(m)}$$

We can rewrite (2.185) in the form

$$A(m_{ML}) = \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{ML} + \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{ML}$$

Here $A(m)$ can be inverted unmerically.

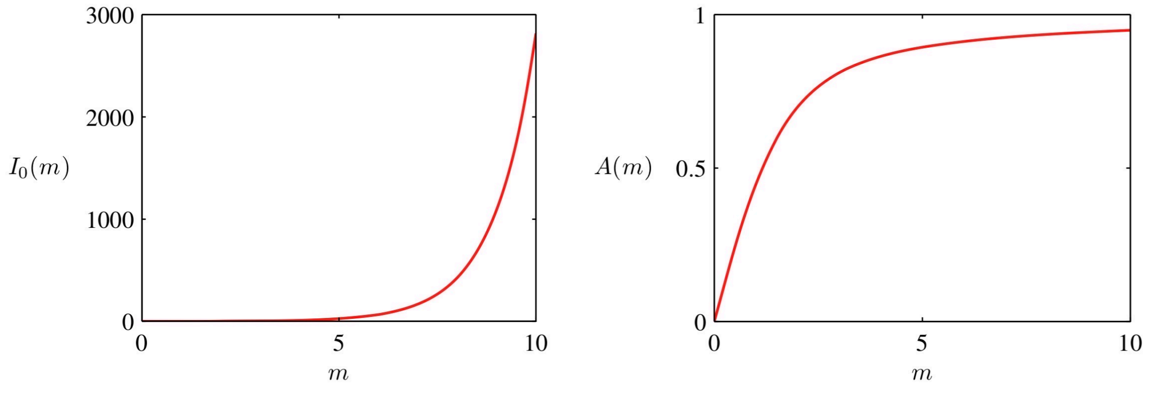
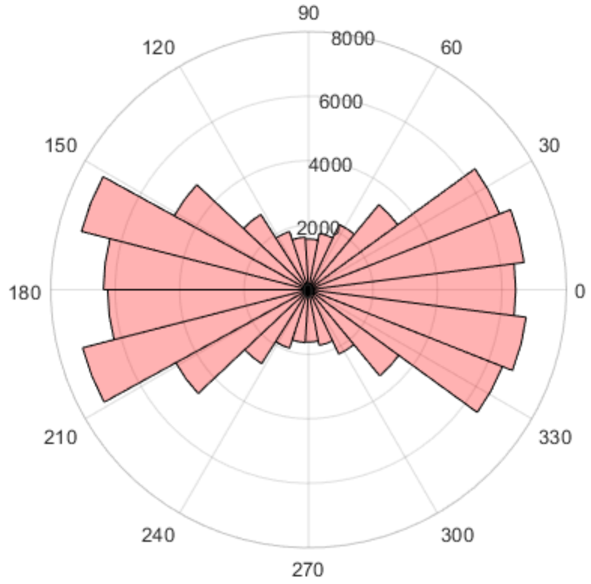


Figure 2.20 Plot of the Bessel function $I_0(m)$ defined by (2.180), together with the function $A(m)$ defined by (2.186).



Remark : other techniques to construct periodic variable

- Histogram in polar coordinates

- Mixtures of von Mises distributions

2.3.9 Mixtures of Gaussians

Limitations of a Gaussian (unimodal)

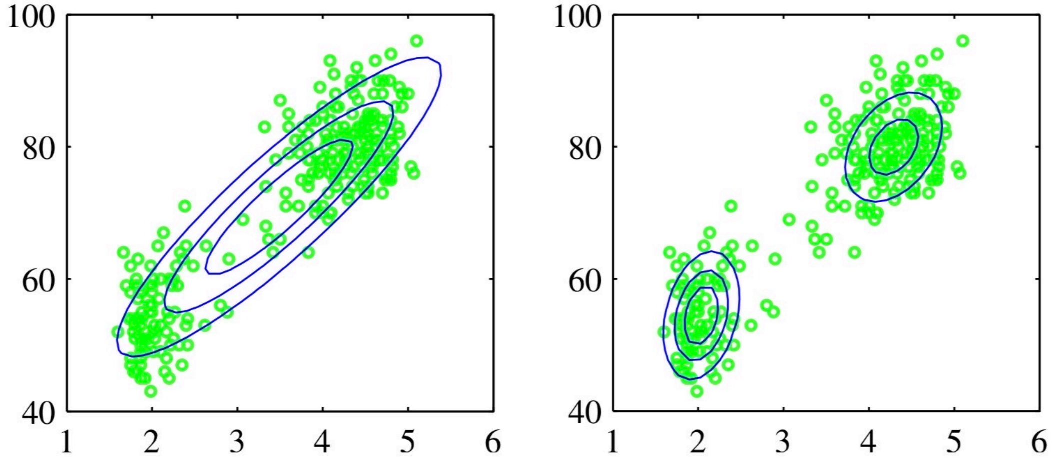


Figure 2.21 Plots of the 'old faithful' data in which the blue curves show contours of constant probability density. On the left is a single Gaussian distribution which has been fitted to the data using maximum likelihood. Note that this distribution fails to capture the two clumps in the data and indeed places much of its probability mass in the central region between the clumps where the data are relatively sparse. On the right the distribution is given by a linear combination of two Gaussians which has been fitted to the data by maximum likelihood using techniques discussed Chapter 9, and which gives a better representation of the data.

Mixture distribution : linear combinations of basic distributions.

Mixture of Gaussians : superposition of K Gaussians

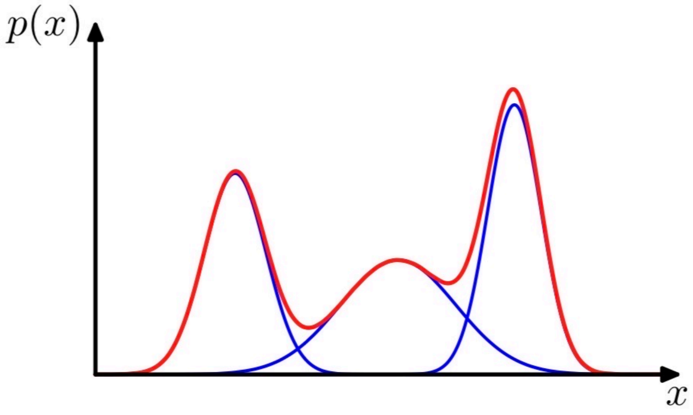
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$$

Each $\mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$ is called a component of mixture

The parameter π_k are called mixing coefficients and satisfies

$$\sum_{k=1}^{K} \pi_k = 1 \quad , \quad 0 \leq \pi_k \leq 1$$

Figure 2.22 Example of a Gaussian mixture distribution in one dimension showing three Gaussians (each scaled by a coefficient) in blue and their sum in red.



$p(\mathbf{x})$ can be rewrite in the form

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

$\pi_k = p(k)$ prior probability of picking the k^{th} component

$\mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) = p(\mathbf{x} | k)$ probability of $\mathbf{x}$ conditioned on k.

Consider the posterior $p(k|\mathbf{x})$ a.k.a. responsibilities

$$\gamma_k(\mathbf{x}) := p(k|\mathbf{x}) = \frac{p(k,\mathbf{x})}{p(\mathbf{x})} = \frac{p(k)\,p(\mathbf{x}|k)}{\sum_\ell p(\ell)\,p(\mathbf{x}|\ell)}$$

$$= \frac{\pi_k\,\mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_\ell \pi_\ell\,\mathcal{N}(\mathbf{x}|\mu_\ell,\Sigma_\ell)}$$

Gaussian mixture is governed by $\pi := \{\pi_1 \dots \pi_K\}$, $\mu := \{\mu_1 \dots \mu_K\}$ and $\Sigma := \{\Sigma_1, \dots \Sigma_K\}$

One way to set these parameters is to use maximum likelihood.

$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\}$$

Maximum likelihood solution for the parameters no longer has a closed-form analytical solution.

Expectation maximization (chapter 9)

2.4 The exponential family

Broad class of distributions called the exponentials family

Random vector $\mathbf{x}$, parameters $\boldsymbol{\eta}$ (called natural parameters)

$$p(\mathbf{x} \mid \boldsymbol{\eta}) := \underline{h(\mathbf{x})}\, g(\boldsymbol{\eta})\, \underline{\exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}} \qquad (2.194)$$

Here $\mathbf{u}(\mathbf{x})$ is some function of $\mathbf{x}$ and $g(\boldsymbol{\eta})$ can be interpreted as the normalization coefficient. i.e.

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}\, d\mathbf{x} = 1 \qquad (2.195)$$

Recall Bernoulli distribution

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \qquad x = 0, 1$$

$$= \exp\{ x \ln \mu + (1-x) \ln (1-\mu) \}$$

$$= (1-\mu) \exp\{ \ln \left(\frac{\mu}{1-\mu} \right) x \}$$

Comparison with (2.194)

$$\eta = \ln \left(\frac{\mu}{1-\mu} \right)$$

Solve for μ to give

$$\mu = \sigma(\eta) := \frac{1}{1+\exp(-\eta)}$$ logistic sigmoid

Thus, Bernoulli distribution can be rewrited in the form

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

we have used $\sigma(-\eta) = 1 - \sigma(\eta)$. Comparison with (2.194)

$$u(x) = x$$

$$h(x) = 1$$

$$g(\eta) = \sigma(-\eta)$$

Next consider the multinomial distribution

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{ \sum_{k=1}^{M} x_k \ln \mu_k \right\}$$

where $\mathbf{x} = (x_1, \ldots x_M)^T$ (one - hot vector)

The standard representation (2.149) so that

$$p(\mathbf{x} \mid \boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^T \mathbf{x})$$

where $\boldsymbol{\eta} = (\eta_1 \ldots \eta_M)^T$ with $\eta_k = \ln \mu_k$. i.e.

$$u(\mathbf{x}) = \mathbf{x}, \qquad h(\mathbf{x}) = 1, \qquad g(\boldsymbol{\eta}) = 1$$

Since $\sum_{k=1}^{M} \mu_k = 1$, parameters η_k are not independent.

i.e. $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. left $M-1$ parameters

$$0 \le \mu_k \le 1 , \quad \sum_{k=1}^{M-1} \mu_k \le 1$$

So the multinomial distribution becomes

$$\exp\left\{ \sum_{k=1}^{M} x_k \ln \mu_k \right\} = \exp\left\{ \sum_{k=1}^{M-1} x_k \ln \mu_k + \overset{1-\sum_{k=1}^{M-1} x_k}{\overset{/\!/}{x_M}} \underbrace{\ln \mu_M}_{\overset{/\!/}{\ln\left(1-\sum_{k=1}^{M-1} \mu_k\right)}} \right\}$$

$$= \exp\left\{ \sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1-\sum_{j=1}^{M-1} \mu_j} \right) + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k \right) \right\}$$

Now identify

$$\ln\left(\frac{\mu_k}{1-\sum_j \mu_j}\right) = \eta_k$$

and

$$\mu_k = \frac{\exp(\eta_j)}{1+\sum_j \exp(\eta_j)}$$

softmax
normalized exponential

In this representation, multinomial distribution

$$p(\boldsymbol{x}\mid\boldsymbol{\eta}) = \left(1+\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\boldsymbol{\eta}^T\boldsymbol{x}), \qquad \boldsymbol{x}\in\mathbb{R}^{M-1}$$

$\boldsymbol{\eta} = (\eta_1 \dots \eta_{M-1})^T.$

$$\mathbf{u}(\mathbf{x}) = \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{M-1} \end{pmatrix} \qquad h(\mathbf{x}) = 1, \qquad g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1}$$

Finally, consider the univariate Gaussian distribution.

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} (x-\mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right\} \end{aligned}$$

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}, \qquad \mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \qquad h(x) = (2\pi)^{-1/2}, \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2} \right)$$

2.4.1 Maximum likelihood and sufficient statistics

Consider the exponential family of distributions over $\mathbf{x}$

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})\, g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \qquad (2.194)$$

Taking the gradient of both side of

① $$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}\, d\mathbf{x} = 1 \qquad (2.195)$$

w.r.t $\boldsymbol{\eta}$, we have

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}\, d\mathbf{x}$$

$$+ g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}\, \mathbf{u}(\mathbf{x})\, d\mathbf{x} = 0$$

Using (2.195) then

$$-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}\,\mathbf{u}(\mathbf{x})\,d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

We therefore obtain the result

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Now consider iid samples denoted by $\mathbf{X} = \{\mathbf{x}_1 \ldots \mathbf{x}_N\}$

for which likelihood

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}$$

Setting the gradient of $\ln p(\mathbf{X}|\eta)$ w.r.t η to zero.

We get the following condition to be satisfied by η_{ML}

$$-\nabla \ln g(\eta_{ML}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

Note that MLE depends on the data only through $\sum_n \mathbf{u}(\mathbf{x}_n)$

called sufficient statistic of (2.194)

Do not need to store the entire data.

E.g. Bernoulli $u(x) = x$. sum of $\{x_n\}$

Gaussian $\mathbf{u}(x) = (x, x^2)^T$. sum of $\{x_n\}$ and $\{x_n^2\}$.

2.4.2 Conjugate prior

For a given prob. density $p(\mathbf{x} | \boldsymbol{\eta})$, seek a prior $p(\boldsymbol{\eta})$ that is conjugate to the likelihood function.

(the posterior has the same functional form as the prior)

For exponential family (2.194), ∃ conjugate prior of $\boldsymbol{\eta}$

$$p(\boldsymbol{\eta} | \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\, g(\boldsymbol{\eta})^{\nu} \exp\{\nu \boldsymbol{\eta}^{T} \boldsymbol{\chi}\}$$

where $f(\boldsymbol{\chi}, \nu)$ is a normalization coefficient and $g(\boldsymbol{\eta})$ is the same function in (2.194)

The posterior

$$p(\eta \mid \mathbf{X}, \chi, \nu) \propto p(\mathbf{X} \mid \eta) \cdot p(\eta \mid \chi, \nu)$$

$$\propto g(\eta)^{\nu+N} \exp\left\{ \eta^T \left(\sum_{n=1}^{N} u(\mathbf{x}_n) + \nu \chi \right) \right\}$$

2.4.3 Noinformative priors

Intend to have as litte influence on the posterior as possible.

Let density or likelihood is given by $p(x|\lambda)$.

Consider noninformative prior $p(\lambda)$

First, $p(\lambda) = \text{constant}$

- If the domain of λ is unbounded, prior cannot be normalized. Such prior is called improper.
- Transformation behavior of density under a nonlinear change of variables

Example 1.

Density of x takes the form

$$p(x \mid \mu) = f(x-\mu)$$

μ is known as location parameter. E.g. $\mathcal{N}(x \mid \mu, \sigma^2)$

Translation invariance

If $x \to \hat{x} := x + c$, then

$$\hat{p}(\hat{x} \mid \hat{\mu}) = f(\hat{x} - \hat{\mu})$$

where we have defined $\hat{\mu} := \mu + c$

Thus $p(x|\mu) = p(\hat{x}|\hat{\mu})$ so density is independent of origin.

Prior distribution should satisfy this translation invariance property.

$$\Rightarrow \quad \int_A^B p(\mu)\, d\mu = \int_{A-C}^{B-C} p(\mu)\, d\mu = \int_A^B p(\mu - c)\, d\mu \qquad \forall A, B$$

So we have $p(\mu - c) = p(\mu)$

Example of location parameter: mean of a Gaussian

The conjugate prior for μ is again Gaussian $p(\mu | \mu_0, \sigma_0^2)$

and we obtain noninformative prior by taking $\sigma_0^2 \longrightarrow \infty$.

Example 2.

Density of x takes the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \qquad \sigma > 0$$

σ is known as scale parameter. E.g. $\mathcal{N}(x|\mu, \sigma^2)$

Scale invariance

If $x \to \hat{x} := cx$, then

$$\hat{p}(\hat{x}|\hat{\sigma}) = \frac{1}{\hat{\sigma}} f\left(\frac{\hat{x}}{\hat{\sigma}}\right)$$

where we have defined $\hat{\sigma} = c\sigma$

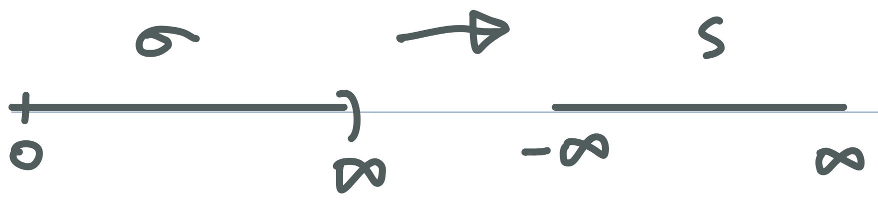
So this transformation corresponds to a change of scale.

Prior distribution should satisfy this scale invariance property.

$$\Rightarrow \quad \int_A^B p(\sigma)\,d\sigma = \int_{A/c}^{B/c} p(\sigma)\,d\sigma = \int_A^B p\left(\tfrac{1}{c}\sigma\right)\tfrac{1}{c}\,d\sigma \quad \forall\, A, B$$

So we have $p(\sigma) = p\left(\frac{1}{c}\sigma\right)\frac{1}{c}$ and hence $p(\sigma) \propto \frac{1}{\sigma}$

Note that this is an improper prior because of $0 < \sigma < \infty$



$$P_s(\ln\sigma) = \underbrace{p(\sigma)}_{\propto \frac{1}{\sigma}} \cdot \left|\frac{d\sigma}{ds}\right| = \text{constant}$$

$s(\sigma) = \ln\sigma \qquad ds = \frac{1}{\sigma}\,d\sigma$

Convenient to think of prior for scale parameter in terms of the density of the log of the parameter.

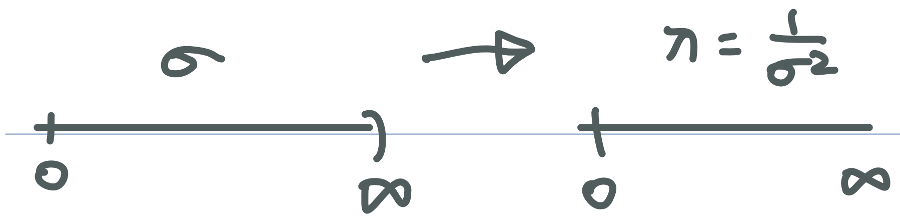
Example of scale parameter: standard deviation σ of a Gaussian

$$\mathcal{N}(x \mid \mu, \sigma^2) \propto \sigma^{-1} \exp\left\{ -(\tilde{x}/\sigma)^2 \right\}$$

where $\tilde{x} := x - \mu$

More convenient to work in terms of the precision

$\lambda = 1/\sigma^2$ rather than σ itself $\qquad d\lambda = \frac{1}{\sigma^3} d\sigma$



$$p_\lambda(\lambda) = \overset{\propto \frac{1}{\sigma}}{p(\sigma)} \cdot \left| \frac{d\sigma}{d\lambda} \right| \propto 1/\lambda$$

We have seen the conjugate prior for λ was $\mathrm{Gam}(\lambda | a_0, b_0)$

The noninformative is obtainet as the special case $a_0 = b_0 = 0$.

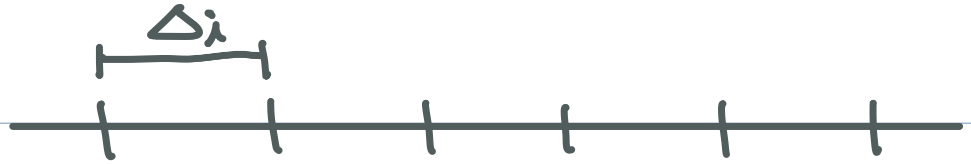
2.5 Nonparametric Methods

Approaches to density modeling

Parametric vs Non parametric (few assumptions)

Histogram method for density estimation

Single continuous random variable x. Partition x into distinct bins of width Δ_i and then count the number n_i of observations of x falling in bin i



We obtain probability values for each bin given by

$$p_i = \frac{n_i}{N \Delta_i}$$

where N is # of total observations. So a model for the density $p(x)$ is piecewise constant over the width Δ_i of each bin, and often the bins are chosen to have the same width $\Delta_i = \Delta$

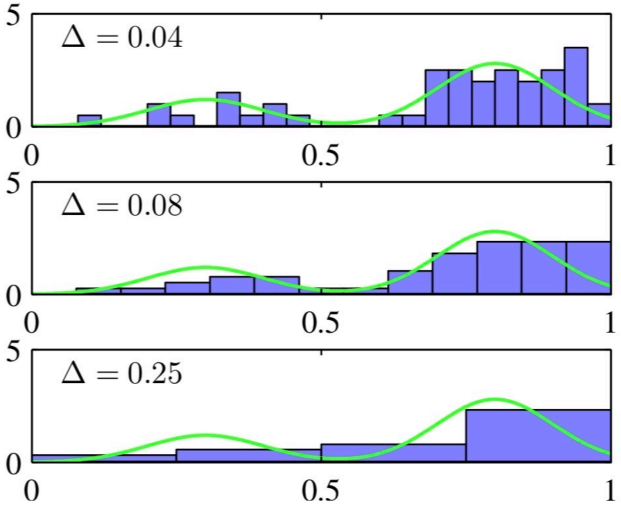


Figure 2.24 An illustration of the histogram approach to density estimation, in which a data set of 50 data points is generated from the distribution shown by the green curve. Histogram density estimates, based on (2.241), with a common bin width Δ are shown for various values of Δ.

Remark

- Effect of a choice of width Δ (smoothing parameter)
- After compting histogram, the data set can be discarded.
- Useful tool for a quick visualization of 1-d or 2-d data
- Limitation of high dimensional data M^D, M bins in D-dim

2.5.1 kernel density estimator

Estimate unknown probability density $p(\mathbf{x})$ in D-dim space.

Consider some small region $\mathcal{R}$ containing $\mathbf{x}$.

Then the probability mass associated with this R

$$P = \int_R p(\mathbf{x})\, d\mathbf{x} \qquad \text{(true prob.)}$$

Suppose we observed N data set drawn from $p(\mathbf{x})$.

Since each point has a probability P of falling within R

$$\mathrm{Bin}(K \mid N, P) = \binom{N}{k} P^k (1-P)^{N-k} \qquad k=0,1,\dots N$$

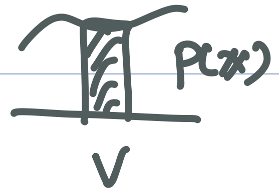
$$\Rightarrow \quad \mathbb{E}[K/N] = P, \qquad \mathrm{var}[K/N] = P(1-P)/N$$

For large N,

$$K \simeq NP$$

If the region R is sufficiently small that $p(\mathbf{x})$ is roughly constant over R, then

$$P \simeq p(\mathbf{x})\, V$$



where V is the volume of R.

Combining these expressions we obtain the density estimate

$$p(x) = \frac{K}{NV} \qquad (2.246)$$

R 영역

Remark: two contradictory assumptions on R and K

We can exploit (2.246) in two different ways

- kernel density estimator (fix V)
- K nearest neighbour method (fix K)

Kernel method in detail (fix R and V)

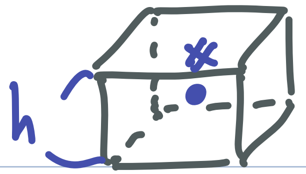
Take the region R to be a small hypercube centered on $\mathbf{x}$

To count the number K of points falling within R, define

$$k(\mathbf{u}) := \begin{cases} 1 & |u_i| \le \frac{1}{2} \qquad i=1, \dots D \\ 0 & \text{otherwise} \end{cases}$$

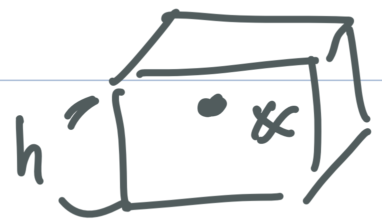
$k(\mathbf{u})$ is an example of kernel function. i.e. the quantity

$k((\mathbf{x} - \mathbf{x}_n)/h) = 1$ if $\mathbf{x}_n$ lies in a cube of side h centered on $\mathbf{x}$ otherwise 0.



The total number of points lying in the cube

$$K := \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$



Substituting this expression into (2.246),

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

where we have used $V = h^D$. (example of kernel density estimator)

We can obtain a smoother density model (Gaussian kernel)

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \exp\left\{ -\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h} \right\}$$

where h represents the standard deviation of Gaussian component and plays the role of a smoothing parameter.

Generally, we can choose any other kernel function $K(\mathbf{u})$ subject to

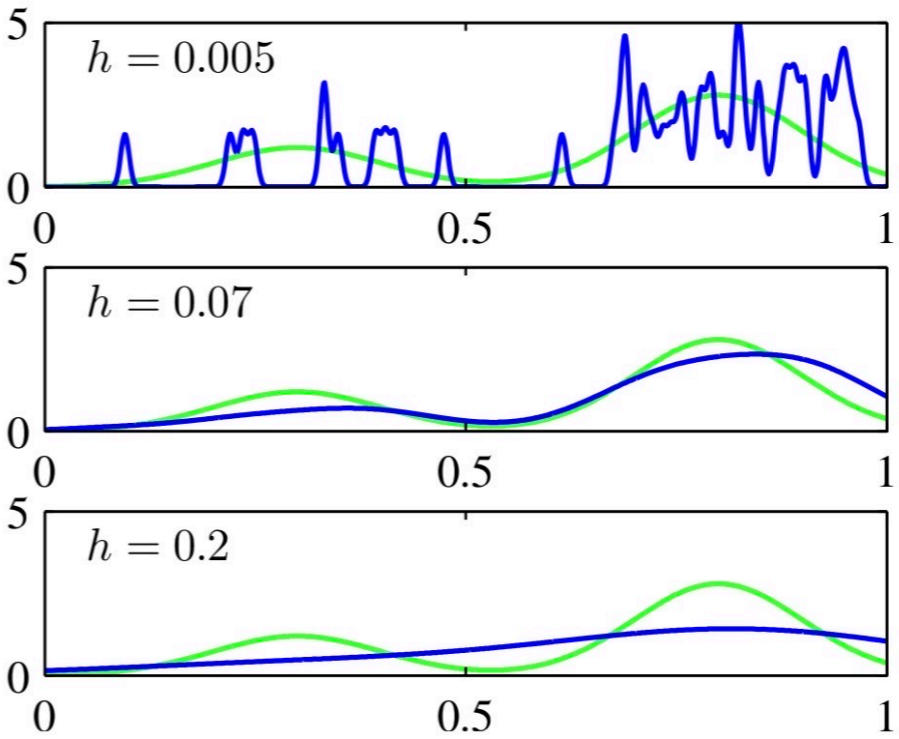
$$K(\mathbf{u}) \geq 0$$

$$\int K(\mathbf{u})\, d\mathbf{u} = 1$$

Remark

- No computation involved in the training phase
- Computational cost grows linearly with the data size.

Figure 2.25 Illustration of the kernel density model (2.250) applied to the same data set used to demonstrate the histogram approach in Figure 2.24. We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model, whereas if it is set too large (bottom panel), then the bimodal nature of the underlying distribution from which the data is generated (shown by the green curve) is washed out. The best density model is obtained for some intermediate value of h (middle panel).



2.5.2 Nearest - neighbour methods

K nearest neighbours

For local density estimation, fix value of K and use the data to find an appropriate value for V.

Consider a sphere centered on $\mathbf{x}$ and allow the radius to grow until it contains K data points. i.e. the radius is not determined (fixed)

The value of K governs the degree of smoothing.

Use (2.246) $p(\mathbf{x}) = \dfrac{K}{NV}$ with KNN method for density estimation

KNN method can be extended to classification.

Apply KNN to each class separately and then make use of Baye's theorem.

N : total $\#$ of data set, N_k : $\#$ of points in C_k

i.e. $\sum_{k=1}^{K} N_k = N$

New point x (fixed). Draw a sphere centered on x containing precisely K points irrespective of their class.

This sphere has the volume V and contains K_k points from the class C_k.

$$p(\mathbf{x} \mid C_k) = \frac{\boxed{K_k}}{\underline{N_k}\,\boxed{V}}$$

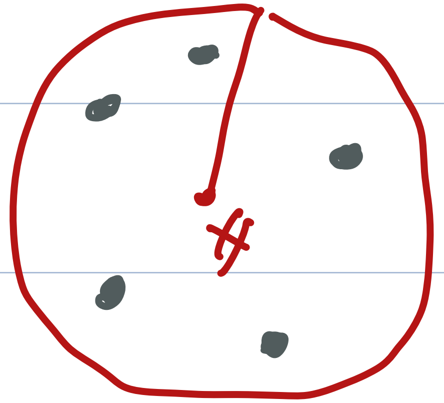
Similarly, the unconditioned density is given by

$$p(\mathbf{x}) = \frac{\underline{K}}{\underline{N}\,\boxed{V}}$$

and class prior $p(C_k) = N_k / N$

Combining these equations and using Baye's theorem

$$p(C_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K}$$



To minimize the probability of misclassification, assign x to the class having the largest posterior probability K_k/K

The particular case of $K=1$ is called nearest-neighbour

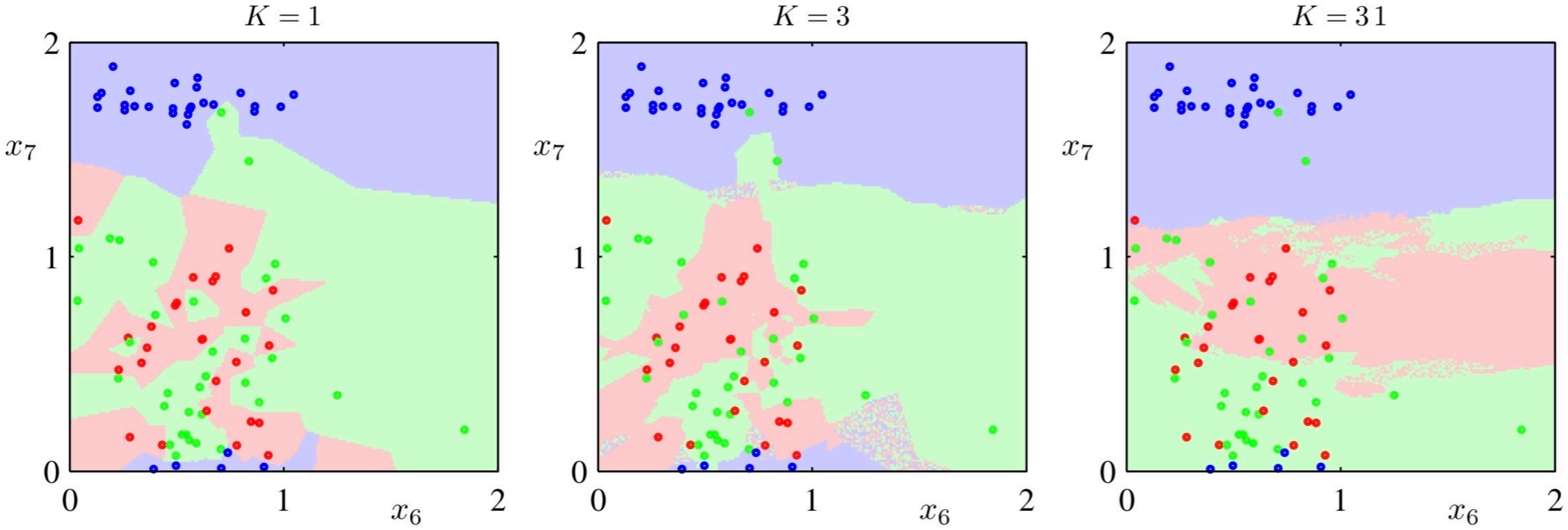


Figure 2.28 Plot of 200 data points from the oil data set showing values of x_6 plotted against x_7, where the red, green, and blue points correspond to the 'laminar', 'annular', and 'homogeneous' classes, respectively. Also shown are the classifications of the input space given by the K-nearest-neighbour algorithm for various values of K.

What happen if $K = N$.

Remark

- k controls the degree of smoothing
- KNN and kernel density methods require the entire data set to stored