Probability Distributions

张倍思

2019/10/14

Outline

1. 2.1 Binary Variables
   - 2.1.1 The beta distribution
2. 2.2 Multinomial Variables
   - 2.2.1 The Dirichlet distribution
3. 2.3 The Gaussian Distribution
   - 2.3.1 Conditional Gaussian distributions
   - 2.3.2 Marginal Gaussian distributions
   - 2.3.3 Bayes'theorem for Gaussian variables
   - 2.3.4 Maximum likelihood for the Gaussian
   - 2.3.5 Sequential estimation

Density estimation

- One role for the distributions discussed in this chapter is to model the probability distribution $p(x)$ of a random variable x, given a finite set $x_1, \ldots, x_N$ of observations.
- This problem is known as *density estimation*

Density estimation

- We begin by considering the binomial and multinomial distributions for discrete random variables and the Gaussian distribution for continuous random variables. These are specific examples of parametric distributions, so-called because they are governed by a small number of adaptive parameters, such as the mean and variance in the case of a Gaussian for example.

2.1 Binary Variables

- Consider a single binary random variable$x \in \{0,1\}$
- x might describe the outcome of flipping a coin, with $x = 1$ representing ‘heads’, and $x = 0$ representing ‘tails’.

$$p(x = 1) = \mu \tag{2.1}$$

where$(0 \le \mu \le 1)$, from which it follows that $p(x = 0) = 1 - \mu$

Bernoulli distribution

The probability distribution over x can therefore be written in the form

$$\mathrm{Bern}(x|\mu) = \mu^{x}(1-\mu)^{1-x} \tag{2.2}$$

which is known as the *Bernoulli distribution*.

Bernoulli distribution

The probability distribution over x can therefore be written in the form

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{2.2}$$

which is known as the Bernoulli distribution.

$$\mathbb{E}[x] = \sum xP(x) = 0 * (1-\mu) + 1 * \mu = \mu \tag{2.3}$$

$$\text{var}[x] = \mathbb{E}[x - \mathbb{E}[x]^2] = \mu(1-\mu) \tag{2.4}$$

Maximum Likelihood Estimation of Bernoulli Distribution

Suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of observed values of x . Under the assumption of independent and identical distribution, the likelihood function is

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

Maximum Likelihood Estimation of Bernoulli Distribution

The log likelihood function is given by

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\} \quad (2.6)$$

The log likelihood function depends on the observations x_n only through $\sum_{n=1}^{N} x_n$,This sum provides an example of a sufficient statistic for the data under this distribution.

Maximum Likelihood Estimation of Bernoulli Distribution

If we set the derivative of $\ln p(\mathcal{D}|\mu)$ with respect to μ equal to zero, we obtain the maximum likelihood estimator.

$$\frac{\partial}{\partial\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N}\left\{\frac{x_n}{\mu} - \frac{1-x_n}{1-\mu}\right\} = 0$$

Maximum Likelihood Estimation of Bernoulli Distribution

If we set the derivative of $\ln p(\mathcal{D}|\mu)$ with respect to μ equal to zero, we obtain the maximum likelihood estimator.

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2.7}$$

which is also known as the *sample mean*.If we denote the number of observations of $x = 1$ (heads) within this data set by m,

$$\mu_{ML} = \frac{m}{N} \tag{2.8}$$

Maximum Likelihood Estimation of Bernoulli Distribution

If $\mathcal{D} = \{1, 1, 1\}, \mu_{ML} = 1$.In this case, the maximum likelihood result would predict that all future observations should give heads.

binomial distribution

We can also work out the distribution of the number m of observations of $x = 1$, given that the data set has size N. This is called the binomial distribution.

$$\text{Bin}(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \tag{2.9}$$

where

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{2.10}$$

binomial distribution

Histogram plot of the binomial distribution (2.9) as a function of m for $N = 10$ and $\mu = 0.25$.

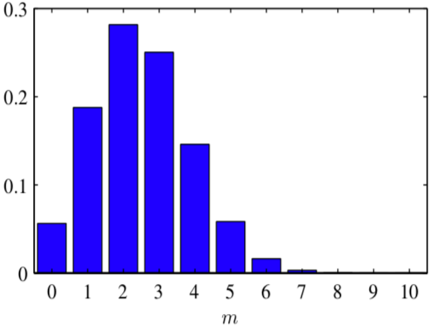


图: 2.1

binomial distribution

We can also work out the distribution of the number m of observations

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.9}$$

where

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{2.10}$$

in case of the binomial theorem :

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

Verify that it is a probability distribution.

binomial distribution

For independent events the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances.

$$\mathbb{E}[m] = \sum_{m=0}^{N} m\mathsf{Bin}(m|N,\mu) = N\mu \tag{2.11}$$

$$\mathsf{var}[m] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2\mathsf{Bin}(m|N,\mu) = N\mu(1-\mu) \tag{2.12}$$

The beta distribution

- As we have already noted, this can give severely overfitted results for small data sets.In order to develop a Bayesian treatment for this problem, we need to introduce a prior distribution $p(\mu)$ over the parameter μ.
- we note that the likelihood function takes the form of the product of factors of the form $\mu^x(1-\mu)^{1-x}$. If we choose a prior to be proportional to powers of μ and $1-\mu$, then the posterior distribution, which is proportional to the product of the prior and the likelihood function, will have the same functional form as the prior. This property is called *conjugacy*

The beta distribution

We therefore choose a prior, called the beta distribution, given by

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.13}$$

where

$$\begin{aligned}\Gamma(x) &= \int_0^\infty u^{x-1}e^{-u}du \\ \Gamma(x+1) &= \int_0^\infty u^{x}e^{-u}du \\ &= x\Gamma(x) \\ \Gamma(1) &= \int_0^\infty e^{-u}du = \left[-e^{-u}\right]_0^\infty = 1\end{aligned}$$

Q:How to verify $\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1$?

The beta distribution

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The beta distribution

The mean and variance of the beta distribution are given by:

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.15}$$

$$\mathsf{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.16}$$

The parameters a and b are often called hyperparameters because they control the distribution of the parameter μ.

The beta distribution

Figure 2.2 shows plots of the beta distribution for various values of the hyperparameters.

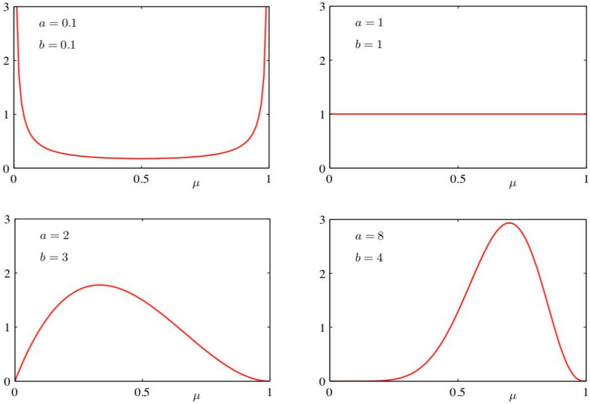


图: 2.2

The beta distribution

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.13}$$

$$\text{Bin}(m \mid N,\mu) = \binom{N}{m}\mu^{m}(1-\mu)^{N-m} \tag{2.9}$$

The posterior distribution of μ is now obtained by multiplying the beta prior (2.13) by the binomial likelihood function (2.9) and normalizing

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.17}$$

where $l = N - m$, and therefore corresponds to the number of ‘tails’in the coin example.

The beta distribution

Its normalization coefficient can therefore be obtained by comparison with (2.13) to give

$$p(\mu|m,l,a,b) \sim \text{Beta}(\mu \mid a+m, b+l) \tag{2.18}$$

hyperparameters a and b in the prior as an effective number of observations of $x=1$ and $x=0$

The beta distribution

Furthermore, the posterior distribution can act as the prior if we subsequently observe additional data.
Figure 2.3 Illustration of one step of sequential Bayesian inference.

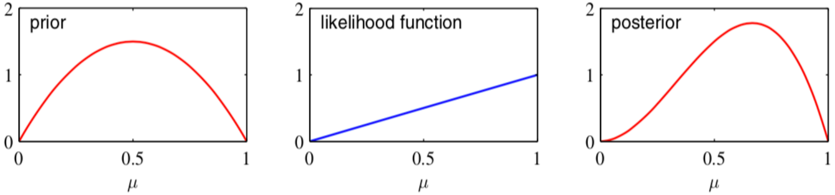


图: 2.3

The beta distribution

If our goal is to predict, as best we can, the outcome of the next trial, then we must evaluate the predictive distribution of x, given the observed data set $\mathcal{D}$.

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}] \tag{2.19}$$

Using the result for the posterior distribution $p(\mu|\mathcal{D})$, we obtain

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b} = \frac{m+a}{N+a+b} \tag{2.20}$$

$m, l \to \infty$

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

The beta distribution

From Figure 2.2, we see that as the number of observations increases, so the posterior distribution becomes more sharply peaked.

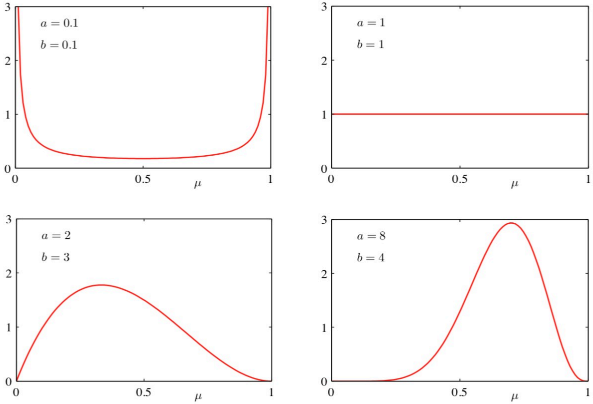


图: 2.2

The beta distribution

- As we observe more and more data, the uncertainty represented by the posterior distribution will steadily decrease.
- Consider a general Bayesian inference problem for a parameter θ for which we have observed a data set $\mathcal{D}$, described by the joint distribution $p(\mu|\mathcal{D})$.

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]] \tag{2.21}$$

$$\mathsf{var}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathsf{var}_\theta[\theta|\mathcal{D}]] + \mathsf{var}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]] \tag{2.24}$$

Multinomial Variables

- For instance if we have a variable that can take K = 6 states
- the state where $x_k = 1$, then $\mathbf{x}$ will be represented by$(x_1, \ldots, x_K)$
- Note that such vectors satisfy $\sum_{k=1}^{K} x_k = 1$
- A particular observation of the variable happens to correspond to the state where $x_3 = 1$, then $\mathbf{x}$ will be represented by

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathsf{T}} \tag{2.25}$$

Multinomial Variables

If we denote the probability of $x_k = 1$ by the parameter μ_k, then the distribution of $\mathbf{x}$ is given

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{2.26}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^\mathsf{T}$,and $\mu_k \geq 0, \sum_k \mu_k = 1$.

Multinomial Variables

It is easily seen that the distribution is normalized

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1 \tag{2.27}$$

and that

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}) = (\mu_1, \ldots, \mu_K)^{\mathsf{T}} = \boldsymbol{\mu} \tag{2.28}$$

Maximum likelihood

Consider a data set $\mathcal{D}$ of N independent observations $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$. The corresponding likelihood function takes the form

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{2.29}$$

$$m_k = \sum_n x_{nk} \tag{2.30}$$

which represent the number of observations of $x_k = 1$. These are called *the sufficient statistics for this distribution*.

Maximum likelihood

This can be achieved using a Lagrange multiplier λ and maximizing:

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1) \tag{2.31}$$

Setting the derivative of (2.31) with respect to μ_k to zero, we obtain

$$\mu_k = -\frac{m_k}{\lambda} \tag{2.32}$$

substituting (2.32) into the constraint$\sum_k \mu_k = 1$ to give $\lambda = -N$

$$\mu_k^{\mathsf{ML}} = \frac{m_k}{N} \tag{2.33}$$

which is the fraction of the N observations for which $x_k = 1$.

Maximum likelihood

We can consider the joint distribution of the quantities $m_1, \ldots, m_K$

$$\text{Mult}(m_1, m_2, \ldots, m_k | \mu, N) = \binom{N}{m_1 m_2 \ldots m_k} \prod_{k=1}^{K} \mu_k^{m_k} \tag{2.34}$$

which is known as the multinomial distribution.

$$\binom{N}{m_1 m_2 \ldots m_k} \equiv \frac{N!}{m_1! m_2! \cdots m_K!} \tag{2.35}$$

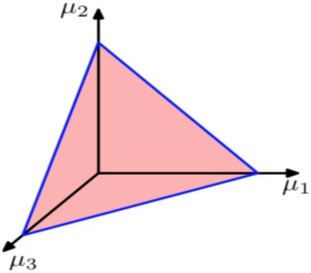
$$N = \sum_{k=1}^{K} m_k \tag{2.36}$$

The Dirichlet distribution

We now introduce a family of prior distributions for the multinomial distribution.

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

where $0 \leq \mu_k \leq 1, \sum_k \mu_k = 1$,$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathsf{T}}$are the parameters of the distribution.



The Dirichlet distribution

We now introduce a family of prior distributions for the multinomial distribution.

$$\mathsf{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.38}$$

where

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k \tag{2.39}$$

which is called the Dirichlet distribution.

The Dirichlet distribution

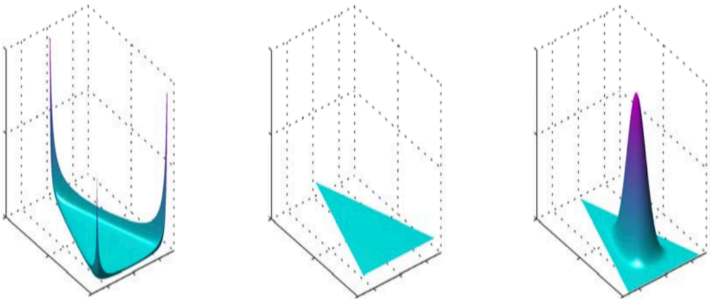


图: 2.5

The Dirichlet distribution

Multiplying the prior (2.38) by the likelihood function (2.34), we obtain the posterior distribution for the parameters μ_k in the form

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.38}$$

$$\text{Mult}(m_1, m_2, \ldots, m_k|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_k} \prod_{k=1}^{K} \mu_k^{m_k} \tag{2.34}$$

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\alpha}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{2.40}$$

The Dirichlet distribution

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\alpha})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \tag{2.40}$$

We see that the posterior distribution again takes the form of a Dirichlet distribution, confirming that the Dirichlet is indeed a conjugate prior for the multinomial.

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) = \mathsf{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\ldots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \tag{2.41}$$

where $\boldsymbol{m} = (m_1,\ldots,m_K)^\mathsf{T}$

The Gaussian Distribution

In the case of a single variable x, the Gaussian distribution can be written in the form

$$\mathcal{N}\left(x \,\middle|\, \mu, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.42}$$

where μ is the mean and σ^2 is the variance.

The Gaussian Distribution

For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \quad (2.43)$$

where $\boldsymbol{\mu}$ is a D-dimensional mean vector and $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix, and$|\boldsymbol{\Sigma}|$denotes the determinant of $\boldsymbol{\Sigma}$.

The Gaussian Distribution

The central limit theorem

$$(x_1 + \cdots + x_N)/N$$

For large N , this distribution tends to a Gaussian, as illustrated in Figure 2.6.

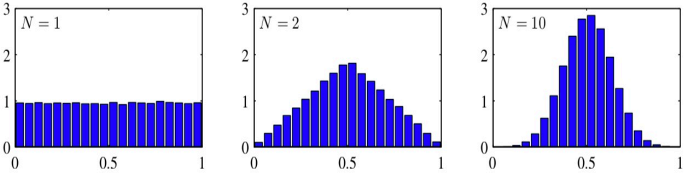


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

The Gaussian Distribution

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \quad (2.43)$$

The functional dependence of the Gaussian on $\mathbf{x}$ is through the quadratic form

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \quad (2.44)$$

The quantity Δ is called the Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$ and reduces to the Euclidean distance when $\boldsymbol{\Sigma}$ is the identity matrix.

The Gaussian Distribution

Now consider the eigenvector equation for the covariance matrix

$$\mathbf{\Sigma}\boldsymbol{u}_i = \lambda_i \mathbf{u}_i \tag{2.45}$$

$$\mathbf{u}_i^\top \mathbf{u}_j = I_{ij} \tag{2.46}$$

$$\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \tag{2.48}$$

$$\mathbf{\Sigma}^{-1} = (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top)^{-1} = (\mathbf{U}^\top)^{-1}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top = \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top \tag{2.49}$$

let $y_i = \mathbf{u}_i^\top(\boldsymbol{x} - \boldsymbol{\mu})$

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \tag{2.50}$$

The Gaussian Distribution

We can interpret y_i as a new coordinate system defined by the orthonormal vectors u_i that are shifted and rotated with respect to the original x_i coordinates.

$$\mathbf{y} = \mathbf{U}^\top(\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.52}$$

where $\mathbf{U}$ is a matrix whose rows are given by $u_i^\top$

The Gaussian Distribution

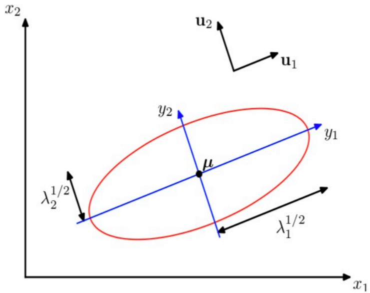


图 2.7: 红色曲线表示二维空间$\boldsymbol{x} = (x_1, x_2)$的高斯分布的常数概率密度的椭圆面，它表示的概率密度为$\exp(-1/2)$，值是在$\boldsymbol{x} = \boldsymbol{\mu}$处计算的。椭圆的轴由协方差矩阵的特征向量$\boldsymbol{u}_i$定义，对应的特征值为$\lambda_i$。

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 $\mathbf{x}$ to the $\mathbf{y}$ coordinate system, we have a Jacobian matrix $\mathbf{J}$ with elements given by

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ij} \tag{2.53}$$

Using the orthonormality property of the matrix $\mathbf{U}$, we see that the square of the determinant of the Jacobian matrix is

$$|\mathbf{J}|^2 = |\mathbf{U}|^2 = |\mathbf{U}| \left|\mathbf{U}^\top\right| = |\mathbf{U}| \left|\mathbf{U}^\top\right| = |\mathbf{I}| = 1 \tag{2.54}$$

as the product of its eigenvalues, and hence

$$|\boldsymbol{\Sigma}| = \prod_{j=1}^{D} \lambda_j \tag{2.55}$$

The Gaussian Distribution

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

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left(-\frac{y_j^2}{2\lambda_j}\right) \tag{2.56}$$

which is the product of D independent univariate Gaussian distributions.

$$\int p(\mathbf{y})d\mathbf{y} = \prod_{j=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left(-\frac{y_j^2}{2\lambda_j}\right) dy_j = 1 \tag{2.57}$$

The Gaussian Distribution

The expectation of $\mathbf{x}$ under the Gaussian distribution is given by

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})d\mathbf{z}\end{aligned} \tag{2.58}$$

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \tag{2.59}$$

The Gaussian Distribution

$$\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{U}\boldsymbol{y} = \sum_{j=1}^{D} y_j \mathbf{u}_j$$

$$\mathbf{z}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{z} = \sum_{k=1}^{D} \frac{y_k^2}{\lambda_k}$$

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^{\top}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^{\top}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\boldsymbol{z}+\boldsymbol{\mu})(\boldsymbol{z}+\boldsymbol{\mu})^{\top}d\mathbf{z}\end{aligned}$$

The Gaussian Distribution

$$
\begin{aligned}
&\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}\mathbf{z}\mathbf{z}^\top d\mathbf{z}\\
=&\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\sum_{i=1}^{D}\sum_{j=1}^{D}\int \exp\left\{-\sum_{k=1}^{D}\frac{y_k^2}{2\lambda_k}\right\}y_i y_j \mathbf{u}_i\mathbf{u}_j^\top d\mathbf{y}
\end{aligned}
$$

$$
\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^\top\lambda_i = \boldsymbol{\Sigma} \tag{2.61}
$$

$$
\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top \tag{2.62}
$$

$$
\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^\top] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top = \boldsymbol{\Sigma} \tag{2.64}
$$

The Gaussian Distribution

- A general symmetric covariance matrix $\mathbf{\Sigma}$ will have $D(D+1)/2$ independent parameters, and there are another Dindependent parameters in , giving $D(D+3)/2$ parameters in total. For large D, the total number of parameters therefore grows quadratically with D, and the computational task of manipulating and inverting large matrices can become prohibitive.
- A further limitation of the Gaussian distribution is that it is intrinsically unimodal (i.e., has a single maximum) and so is unable to provide a good approximation to multimodal distributions.

Conditional Gaussian distributions

- If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.
- Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{2.65}$$

Conditional Gaussian distributions

- If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.
- Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.66}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{2.67}$$

Conditional Gaussian distributions

In many situations, it will be convenient to work with the inverse of the covariance matrix,which is known as the precision matrix.

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \tag{2.68}$$

the partitioned form of the precision matrix

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{2.69}$$

Because the inverse of a symmetric matrix is also symmetric, we see that $\boldsymbol{\Lambda}_{ab}$ and $\boldsymbol{\Lambda}_{bb}$ are symmetric, while $\boldsymbol{\Lambda}_{ab} = \Lambda_{ba}^{\top}$.

Conditional Gaussian distributions

Let us begin by finding an expression for the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$.

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{2.44}$$

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \\
&-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu_a})^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu_a}) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu_a})^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu_b}) \\
&-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu_b})^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu_b})^\top \boldsymbol{\Lambda_{bb}}(\mathbf{x}_b - \boldsymbol{\mu_b}) \qquad (2.70)
\end{aligned}$$

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

Conditional Gaussian distributions

Because this distribution is completely characterized by its mean and its covariance, our goal will be to identify expressions for the mean and covariance

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \qquad (2.71)$$

we can immediately equate the matrix of coefficients entering the second order term in $\mathbf{x}$ to the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and the coefficient of the linear term in $\mathbf{x}$ to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, from which we can obtain $\boldsymbol{\mu}$.

Conditional Gaussian distributions

We will denote the mean and covariance of this distribution by $\boldsymbol{\mu}_{a|b}$and $\boldsymbol{\Sigma}_{a|b}$

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu_a})^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu_a}) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu_a})^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu_b}) \\ -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu_b})^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu_b})^\top \boldsymbol{\Lambda_{bb}}(\mathbf{x}_b - \boldsymbol{\mu_b}) \end{aligned} \qquad (2.70)$$

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

$$-\frac{1}{2}\mathbf{x}_a^\top \boldsymbol{\Lambda}_{aa}\mathbf{x}_a \qquad (2.72)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \qquad (2.73)$$

Conditional Gaussian distributions

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

$$\mathbf{x}_a^\top \{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \tag{2.74}$$

the coefficient of $\mathbf{x}_a$ in this expression must equal $\boldsymbol{\Sigma}_{a|b}^{-1}$ and hence

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

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \qquad (2.75) \end{aligned}$$

Conditional Gaussian distributions

we make use of the following identity for the inverse of a partitioned matrix

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

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \qquad (2.77)$$

Using the definition

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \qquad (2.78)$$

Conditional Gaussian distributions

we make use of the following identity for the inverse of a partitioned matrix

$$\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \tag{2.79}$$

$$\boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \tag{2.80}$$

we obtain the following expressions for the mean and covariance

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \qquad (2.81) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \\ &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \qquad (2.82)\end{aligned}$$

Note that the mean of the conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$, given by (2.81), is a linear function of $\boldsymbol{x}_b$ and that the covariance, given by (2.82), is independent of $\boldsymbol{x}_a$.

Marginal Gaussian distributions

We have seen that if a joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is Gaussian, then the conditional distribution$p(\mathbf{x}_a|\mathbf{x}_b)$ will again be Gaussian. Now we turn to a discussion of the marginal distribution given by

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \tag{2.83}$$

our strategy for evaluating this distribution efficiently will be to focus on the quadratic form in the exponent of the joint distribution and thereby to identify the mean and covariance of the marginal distribution $p(\mathbf{x}_a)$

Marginal Gaussian distributions

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
&-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_{\boldsymbol{a}})^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_{\boldsymbol{a}}) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_{\boldsymbol{a}})^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_{\boldsymbol{b}}) \\
&-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_{\boldsymbol{b}})^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_{\boldsymbol{b}})^\top \boldsymbol{\Lambda}_{\boldsymbol{bb}}(\mathbf{x}_b-\boldsymbol{\mu}_{\boldsymbol{b}}) \qquad \text{(2.70)}
\end{aligned}
$$

Picking out just those terms that involve $\mathbf{x}_b$, we have

$$
-\frac{1}{2}\mathbf{x}_b^\mathsf{T}\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^\top\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m})^\top\boldsymbol{\Lambda}_{bb}^{-1}(\mathbf{x}_b-\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^\top\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m} \tag{2.84}
$$

we have defined

$$
\mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) \tag{2.85}
$$

Marginal Gaussian distributions

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b \tag{2.83}$$

we see that the integration over x_b required by (2.83) will take the form

$$\int \exp\left\{-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m})\right\} d\mathbf{x}_b \tag{2.86}$$

Marginal Gaussian distributions

$$\begin{aligned}&\frac{1}{2}\mathbf{m}^{\top}\mathbf{\Lambda}_{bb}^{-1}\mathbf{m}-\frac{1}{2}\mathbf{x}_a^{\top}\mathbf{\Lambda}_{aa}\mathbf{x}_a+\mathbf{x}_a^{\top}\{\mathbf{\Lambda}_{aa}\mu_a+\mathbf{\Lambda}_{ab}\mu_b\}+\text{const}\\=&-\frac{1}{2}\mathbf{x}_a^{\top}(\mathbf{\Lambda}_{aa}-\mathbf{\Lambda}_{ba}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\mathbf{x}_a+\mathbf{x}_a^{\top}(\mathbf{\Lambda}_{aa}-\mathbf{\Lambda}_{ba}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\boldsymbol{\mu}_a+\text{const}\end{aligned} \tag{2.87}$$

$$\mathbf{\Sigma}_a=(\mathbf{\Lambda}_{aa}-\mathbf{\Lambda}_{ba}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})^{-1} \tag{2.88}$$

$$\mathbf{\Sigma}_a(\mathbf{\Lambda}_{aa}-\mathbf{\Lambda}_{ba}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\boldsymbol{\mu}_a=\boldsymbol{\mu}_a \tag{2.89}$$

Marginal Gaussian distributions

$$\begin{pmatrix} \mathbf{\Lambda}_{aa} & \mathbf{\Lambda}_{ab} \\ \mathbf{\Lambda}_{ba} & \mathbf{\Lambda}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix} \tag{2.90}$$

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{2.76}$$

$$\mathbf{\Sigma}_{aa} = (\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})^{-1} \tag{2.91}$$

$$\mathbb{E}[\mathbf{x}_a] = \boldsymbol{\mu}_a \tag{2.92}$$

$$\mathrm{cov}[\mathbf{x}_a] = \mathbf{\Sigma}_{aa} \tag{2.93}$$

conclusion

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

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.94}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \ \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{2.95}$$

Conditional distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{2.96}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{2.97}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{2.98}$$

Marginal Gaussian distributions

We illustrate the idea of conditional and marginal distributions associated with a multivariate Gaussian using an example involving two variables in Figure 2.9.

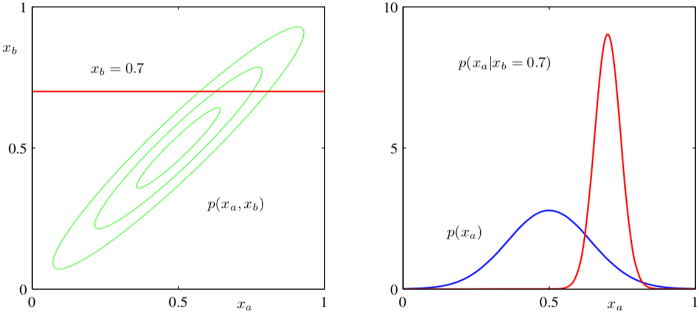


图: 2.9

Bayes'theorem for Gaussian variables

We shall take the marginal and conditional distributions to be

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.99}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \tag{2.100}$$

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

Bayes'theorem for Gaussian variables

First we find an expression for the joint distribution over x and y. To do this, we define

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \tag{2.101}$$

and then consider the log of the joint distribution

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\ &- \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) + \text{const} \end{aligned} \tag{2.102}$$

where 'const'denotes terms independent of $\mathbf{x}$ and $\mathbf{y}$.

Bayes’theorem for Gaussian variables

Because this distribution is completely characterized by its mean and its covariance, our goal will be to identify expressions for the mean and covariance

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\mathrm{const} \qquad (2.71)$$

we can immediately equate the matrix of coefficients entering the second order term in $\mathbf{x}$ to the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and the coefficient of the linear term in $\mathbf{x}$ to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, from which we can obtain $\boldsymbol{\mu}$.

Bayes'theorem for Gaussian variables

To find the precision of this Gaussian

$$\begin{aligned} & -\frac{1}{2}\mathbf{x}^\top(\mathbf{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A})\mathbf{x}-\frac{1}{2}\mathbf{y}^\top\mathbf{L}\mathbf{y}+\frac{1}{2}\mathbf{y}^\top\mathbf{L}\mathbf{A}+\frac{1}{2}\mathbf{x}^\top\mathbf{A}^\top\mathbf{L}\mathbf{y} \\ =& -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^\top\begin{pmatrix}\mathbf{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A} & -\mathbf{A}^\top\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \\ =& -\frac{1}{2}\mathbf{z}^\top\mathbf{R}\mathbf{z} \end{aligned} \tag{2.103}$$

and so the Gaussian distribution over $\mathbf{z}$ has precision (inverse covariance) matrix given by

$$\mathbf{R}=\begin{pmatrix}\mathbf{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A} & -\mathbf{A}^\top\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix} \tag{2.104}$$

Bayes'theorem for Gaussian variables

The covariance matrix is found by taking the inverse of the precision

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \tag{2.76}$$

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \tag{2.77}$$

$$\mathrm{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\top}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top} \end{pmatrix} \tag{2.105}$$

Bayes'theorem for Gaussian variables

Similarly, we can find the mean of the Gaussian distribution over $\mathbf{z}$

$$\mathbf{x}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^{\top}\mathbf{A}^{\top}\mathbf{L}\mathbf{b} + \mathbf{y}^{\top}\mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{\top} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^{\top}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \qquad (2.106)$$

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^{\top}\boldsymbol{L}\boldsymbol{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix} \qquad (2.108)$$

Bayes'theorem for Gaussian variables

$$\mathbf{x}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^{\top}\mathbf{A}^{\top}\mathbf{L}\mathbf{b} + \mathbf{y}^{\top}\mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{\top} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^{\top}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} \tag{2.106}$$

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^{\top}\boldsymbol{L}\boldsymbol{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix} \tag{2.108}$$

Next we find an expression for the marginal distribution $p(y)$ in which we have marginalized over $\mathbf{x}$.

$$\mathbb{E}[\mathbf{y}] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \tag{2.109}$$

$$\operatorname{cov}[\mathbf{y}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top} \tag{2.110}$$

Bayes'theorem for Gaussian variables

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \tag{2.73}$$

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned} \tag{2.75}$$

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}\left\{\mathbf{A}^\top\mathbf{L}(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\} \tag{2.111}$$

$$\mathrm{cov}[\mathbf{x}|\mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1} \tag{2.112}$$

conclusion

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top}) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

Maximum likelihood for the Gaussian

Given a data set $\mathbf{X}$ in which the observations are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood.

$$\ln p(\boldsymbol{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x_n} - \boldsymbol{\mu}) \tag{2.118}$$

$$\sum_{n=1}^{N}\mathbf{x}_n, \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top \tag{2.119}$$

Maximum likelihood for the Gaussian

the derivative of the log likelihood with respect to $\boldsymbol{\mu}$ is given by

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x_n} - \boldsymbol{\mu}) \tag{2.120}$$

setting this derivative to zero, we obtain the solution for the maximum likelihood estimate of the mean given by

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{2.121}$$

Maximum likelihood for the Gaussian

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^\top \tag{2.122}$$

If we evaluate the expectations of the maximum likelihood solutions under the true distribution, we obtain the following results

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu} \tag{2.123}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N}\boldsymbol{\Sigma} \tag{2.124}$$

correct this bias by defining a different estimator

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^\top \tag{2.125}$$

Sequential estimation

Sequential methods allow data points to be processed one at a time and then discarded dissect out the contribution from the final data point $\mathbf{x}_N$, we obtain

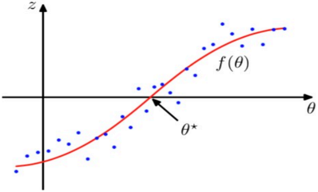
$$
\begin{aligned}
\boldsymbol{\mu}_{ML}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N} - \mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{ML}^{(N-1)} \\
&= \boldsymbol{\mu}_{ML}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{ML}^{(N-1)}) \qquad (2.126)
\end{aligned}
$$

Sequential estimation

we will not always be able to derive a sequential algorithm by this route, and so we seek a more general formulation of sequential learning, which leads us to the Robbins-Monro algorithm. The conditional expectation of z given θ defines a deterministic function $f(\theta)$ that is given by

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)dz \tag{2.127}$$

Functions defined in this way are called regression functions.



Sequential estimation

Our goal is to find the root θ^* at which $f(\theta^*) = 0$.
The following general procedure for solving such problems was given by Robbins and Monro (1951)
We shall assume that the conditional variance of z is finite so that

$$\mathbb{E}[(z - f)^2|\theta] < \infty \tag{2.128}$$

TheRobbins-Monro procedure then defines a sequence of successive estimates of the root given by

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} z(\theta^{(N-1)}) \tag{2.129}$$

where $z(\theta^{(N)})$ is an observed value of Z when θ takes the value $\theta^{(N)}$.

Sequential estimation

The coefficientsa_N represent a sequence of positive numbers that satisfy the conditions

$$\lim_{N\to\infty} a_N = 0 \tag{2.130}$$

$$\sum_{N=1}^{\infty} a_N = \infty \tag{2.131}$$

$$\sum_{N=1}^{\infty} a_N^2 < \infty \tag{2.132}$$

Note that the first condition ensures that the successive corrections decrease in magnitude so that the process can converge to a limiting value. The second condition is required to ensure that the algorithm does not converge short of the root, and the third condition is needed to ensure that the accumulated noise has finite variance and hence does not spoil convergence.

Sequential estimation

let us consider how a general maximum likelihood problem can be solved sequentially using the Robbins-Monro algorithm.

$$\frac{\partial}{\partial\theta}\left\{-\frac{1}{N}\sum_{n=1}^{N}\ln p(x_n|\theta)\right\}\Bigg|_{\theta_{ML}} = 0 \tag{2.133}$$

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta) = \mathbb{E}_x\left[-\frac{\partial}{\partial\theta}\ln p(x|\theta)\right] \tag{2.134}$$

Sequential estimation

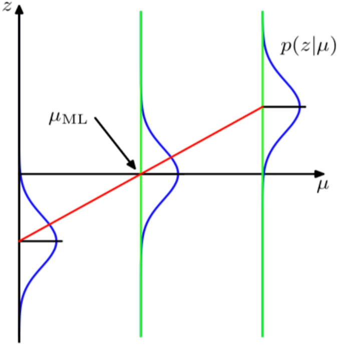
$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \ln p(x_N | \theta^{(N-1)}) \tag{2.135}$$

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

Thus the distribution of z is Gaussian with mean $\mu - \mu_{ML}$,

Sequential estimation

Figure 2.11 In the case of a Gaussian distribution, with θ corresponding to the mean μ, the regression function illustrated in Figure 2.10 takes the form of a straight line, as shown in red. In this case, the random variable z corresponds to the derivative of the log likelihood function and is given by $(x - \mu_{\mathrm{ML}})/\sigma^2$, and its expectation that defines the regression function is a straight line given by $(\mu - \mu_{\mathrm{ML}})/\sigma^2$. The root of the regression function corresponds to the maximum likelihood estimator μ_{ML}.



Thank You