PRML Note
C02 Probability Distributions

Yang Zhao

Department of Automation, Tsinghua University

- Given a finite set $x_1, x_2, \cdots, x_N$ of observation, which is i.i.d, modeling the probability distribution $p(x)$ of a random variable x, which is called density estimation, is ill-posed.
- *conjugate prior*: leads to posterior distributions having the same functional form as the prior.
- *parametric approach*: must assume the specific functional form for the distribution.(Limitation)
- *non-parametric approach*: 1.depends on size of the data set. 2.still has parameters, which control the model complexxity rather than the form of the distribution

1 Binary Variables

- $x \in \{0, 1\}$, let $0 \leq \mu \leq 1$ we have

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

The Bernoulli distribution

$$Bern(x|\mu) = \mu^x(1 - \mu)^{1-x}$$
$$\mathbb{E}[x] = \mu$$
$$var[x] = \mu(1 - \mu)$$

- a data set $D = \{x_1, x_2, \cdots, x_N\}$

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

Let

$$\frac{\partial lnp(D|\mu)}{\partial \mu} = 0$$

we have

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

where m is the number of observations of $x = 1$ within this data.

- *binomial distribution*

$$\begin{aligned} Bin(m|N,\mu) &= C_N^m \mu^m (1-\mu)^{N-m} \\ \mathbb{E}[m] &= N\mu \\ var[m] &= N\mu(1-\mu) \end{aligned}$$

Using the result that for independent variables x and z, $\mathbb{E}[x+z] = \mathbb{E}x + \mathbb{E}z$ and $var[x+z] = varx + varz$, we can easily get the expectation and variance.

- *beta distribution*

$$\begin{aligned} Beta(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} \\ var[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned} \tag{1}$$

where

$$\Gamma(x) = \int_0^\infty \mu^{x-1} e^x dx$$

and

$$\int_0^1 Beta(\mu|a,b) d\mu = 1$$

The parameters a and b are called *hyperparameters*

- posterior distribution

$$\begin{aligned} p(\mu|m,N,a,b) &\propto p(\mu|a,b)p(m|N,\mu) \\ &\propto \mu^{m+a-1}(1-\mu)^{N-m+b-1} \end{aligned}$$

where $l = N - m$. Comparing with the equation 1, we have

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

in which the parameters a, b is the effective number of observations that need not be integers.

- *Sequential approach*: make use of observations once at a time, or in small batches, then discard them before the next observations.

  1. independent of the choice of prior and of likelihood function
  2. depends only on the assumption of i.i.d data

  This method is used in

1. real-time learning scenarios
2. large data sets
3. maximum likelihood methods can also be cast into a sequential framework.

- Given observed data set D, to predict x

$$\begin{aligned} p(x=1|D) &= \int_0^1 p(x=1|\mu)p(\mu|D)d\mu \\ &= \int_0^1 \mu p(\mu|D)d\mu \\ &= \mathbb{E}[\mu|D] \\ &= \frac{m+a}{m+a+l+b} \end{aligned}$$

- A general property of Bayesian learning: As observing more and more data, the uncertainly represented by the posterior distribution will steadily decrease.

- The Bayesian and maximum likelihood results will agree in the limit of an infinitely large data set.
  For a finite data set: posterior mean for μ always lies between the prior mean and the maximum likelihood estimate for μ.

- Consider a general Bayesian inference problem for a parameter θ, observed a data set D, the joint distribution $p(\theta, D)$. We have the following result

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_D[\mathbb{E}_\theta[\theta|D]] \tag{2}$$
$$var_\theta[\theta] = \mathbb{E}_D[var_\theta[\theta|D]] + var_D[\mathbb{E}_\theta[\theta|D]] \tag{3}$$

  The process to get the equation 2 and 3 is below:

$$\begin{aligned} \mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int (\int xp(x|y)dx)p(y)dy \\ &= \iint xp(x|y)p(y)dxdy \\ &= \iint xp(x,y)dxdy \\ &= \int x(\int p(x,y)dy)dx \\ &= \int xp(x)dx \\ &= \mathbb{E}[x] \end{aligned}$$

  which is equal to the equation 2. Let

$$\begin{aligned} a(y) &= \int xp(x|y)dx = \mathbb{E}[x|y] \\ b(y) &= \int x^2 p(x|y)dx = \mathbb{E}[x^2|y] \\ E &= \mathbb{E}_y[\mathbb{E}_x[x|y]] = \mathbb{E}[x] \end{aligned}$$

Thus,

$$\begin{aligned} var_x[x|y] &= \mathbb{E}[x^2|y] - (\mathbb{E}[x|y])^2 \\ &= b(y) - a^2(y) \end{aligned} \tag{4}$$

So

$$\mathbb{E}_y[var_x[x|y]] = \int (b(y) - a^2(y))p(y)dy \tag{5}$$

$$\begin{aligned} var_y[\mathbb{E}_x[x|y]] &= \mathbb{E}\big((\mathbb{E}_x(x|y) - \mathbb{E}_y\mathbb{E}_x(x|y))^2\big) \\ &= \int (a(y) - E)^2 p(y)dy \end{aligned} \tag{6}$$

Add the equation 5 and 6, we have

$$\begin{aligned} &\mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]] \\ &= \int (b(y) - a^2(y))p(y)dy + \int (a(y) - E)^2 p(y)dy \\ &= \int [(a(y) - E)^2 + (b(y) - a^2(y))]p(y)dy \\ &= \int [E^2 + b(y) - 2Ea(y)]p(y)dy \\ &= E^2 \int p(y)dy + \int b(y)p(y)dy - 2E \int a(y)p(y)dy \\ &= E^2 + \mathbb{E}x^2 - 2E^2 \\ &= \mathbb{E}x^2 - E^2 \\ &= var[x] \end{aligned}$$

2 Multinomial Variables

- 1-of-K scheme $\boldsymbol{x} = (0, 0, \cdots, 1, \cdots, 0, 0)^T$
  denote the probability of $x_k = 1$ by the parameter μ_k, then the distribution of $\boldsymbol{x}$ is given

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{7}$$

  where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_K)^T$, $\mu_k \geq 0$ and $\sum_k \mu_k = 1$. This distribution can be regarded as a generalization of the Bernoulli distribution to more than two outcomes.

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

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{\mu}] = \boldsymbol{\mu}$$

- Consider a data set D of N independent observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$, the corresponding likelihood function takes the form

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

where $m_k = \sum_n x_{nk}$, which represent the number of observations of $x_k = 1$. And the maximum likelihood solution is below:

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

- *Multinomial distribution*

$$Mult(m_1, m_2, \cdots, m_k|\boldsymbol{\mu}, N) = \frac{N!}{m_1! m_2! \cdots m_k!} \prod_{k=1}^{K} \mu_k^{m_k}$$

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

- *Dirichlet distribution*

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

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^T$ and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

- posterior distribution

$$\begin{aligned} p(\boldsymbol{\mu}|D, \boldsymbol{\alpha}) &\propto p(D|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \\ &= Dir(\boldsymbol{\mu}|\boldsymbol{\alpha} + \boldsymbol{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \end{aligned}$$

3 The Gaussian Distribution

- *The Gaussian Distribution*

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \tag{9}$$

- For the univariate Gaussian distribution, we have obtain below propositions at Chapter 1

$$\int \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

$$\int \mathcal{N}(x|\mu, \sigma^2) x dx = \mu$$

$$\int \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

- *quadratic form*

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

which is called *Mahalanobis distance* from $\boldsymbol{\mu}$ to $\boldsymbol{x}$

- Without loss of generality, we can assume the matrix $\mathbf{\Sigma}$ to be symmetric. To prove this, we assume the matrix $\mathbf{\Sigma}$ is non-symmetric, we have

$$\begin{aligned}\boldsymbol{x}^T\mathbf{\Sigma}\boldsymbol{x} &= \boldsymbol{x}^T(\frac{\mathbf{\Sigma}+\mathbf{\Sigma}^T}{2}+\frac{\mathbf{\Sigma}-\mathbf{\Sigma}^T}{2})\boldsymbol{x} \\ &= \boldsymbol{x}^T(\frac{\mathbf{\Sigma}+\mathbf{\Sigma}^T}{2})\boldsymbol{x}\end{aligned}$$

  where $(\mathbf{\Sigma}+\mathbf{\Sigma}^T)/2$ is symmetric.

- For $\mathbf{\Sigma}$ is a real, symmetric matrix, let $\boldsymbol{U}=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_D)$ be an orthonormal matrix, i.e. $\boldsymbol{U}\boldsymbol{U}^T=\boldsymbol{I}$, we have

$$\begin{aligned}\mathbf{\Sigma} &= \boldsymbol{U}diag(\lambda_1,\lambda_2,\cdots,\lambda_D)\boldsymbol{U}^T \\ &= (\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_D)\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D\end{pmatrix}\begin{pmatrix}\boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \vdots \\ \boldsymbol{u}_D^T\end{pmatrix}\end{aligned} \tag{10}$$

  where λ_k is the eigenvalue of $\mathbf{\Sigma}$ and $\boldsymbol{u}_k$ is the corresponding eigenvector. And the equation 10 can be rewritten as below

$$\mathbf{\Sigma}=\sum_{i=1}^{D}\lambda_i\boldsymbol{u}_i\boldsymbol{u}_i^T \tag{11}$$

  and

$$\mathbf{\Sigma}^{-1}=\sum_{i=1}^{D}\frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T \tag{12}$$

  if we define that $\boldsymbol{y}=\boldsymbol{U}^T(\boldsymbol{x}-\boldsymbol{\mu})$ and $\boldsymbol{y}=(y_1,y_2,\cdots,y_D)^T$, we have

$$\Delta^2=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \tag{13}$$

- For the Gaussian distribution to be well defined, it is necessary for all of eigenvalues λ_i of the covariance matrix to be strictly positive, otherwise the distribution cannot be properly normalized.

- from x to y, the Jacobian matrix J with elements given by

$$\boldsymbol{J}_{ij}=\frac{\partial x_i}{\partial y_j}=\boldsymbol{U}_{ji}$$

  So $|\boldsymbol{J}|=1$ and $|\Sigma|^{1/2}=\prod_{j=1}^{D}\lambda_j^{1/2}$ Thus in the y coordinate system, the Gaussian distribution takes the form

$$p(\boldsymbol{y})=p(\boldsymbol{x})|\boldsymbol{J}|=\prod_{j=1}^{D}\frac{e^{-y_j^2/2\lambda_j}}{(2\pi\lambda_j)^{1/2}} \tag{14}$$

- For the multivariate Gaussian distribution, we have

$$\int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{x} = 1 \tag{15}$$

$$\mathbb{E}(\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \boldsymbol{x} d\boldsymbol{x} = \boldsymbol{\mu} \tag{16}$$

$$\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) = \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \tag{17}$$

- To prove equation 15

$$\begin{aligned}
\int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{x} &= \int p(\boldsymbol{y}) d\boldsymbol{y} \\
&= \prod_{j=1}^{D} \int \frac{e^{-y_j^2/2\lambda_j}}{(2\pi\lambda_j)^{1/2}} dy_j \\
&= 1
\end{aligned}$$

  which is D production of univariate Gaussian distribution.

- To prove equation 16

$$\begin{aligned}
\mathbb{E}(x) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \boldsymbol{x} d\boldsymbol{x} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} (\boldsymbol{z} + \boldsymbol{\mu}) d\boldsymbol{z}
\end{aligned}$$

  Obviously, $e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}}$ is an even function, which means

$$\int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} \boldsymbol{z} d\boldsymbol{z} = 0 \tag{18}$$

  Thus

$$\mathbb{E}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \boldsymbol{\mu} \int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} d\boldsymbol{z} = \boldsymbol{\mu} \tag{19}$$

- To prove equation 17

$$\begin{aligned}
\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} (\boldsymbol{z} + \boldsymbol{\mu})(\boldsymbol{z} + \boldsymbol{\mu})^T d\boldsymbol{z} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} (\boldsymbol{z}\boldsymbol{z}^T + \boldsymbol{z}\boldsymbol{\mu}^T + \boldsymbol{z}^T\boldsymbol{\mu} + \boldsymbol{\mu}\boldsymbol{\mu}^T) d\boldsymbol{z} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int e^{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}} \boldsymbol{z}\boldsymbol{z}^T d\boldsymbol{z}
\end{aligned} \tag{20}$$

  where using equation 18 to eliminate the cross-terms $\boldsymbol{\mu}\boldsymbol{z}^T$ and $\boldsymbol{z}\boldsymbol{\mu}^T$. The method we introduce below is really important. Because $\boldsymbol{y} = \boldsymbol{U}^T(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{U}^T\boldsymbol{z}$, we get

$\boldsymbol{z} = \boldsymbol{U}\boldsymbol{y} = \sum_{j=1}^{D} y_j \boldsymbol{u}_j$, which gives

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

where the integral will vanish by symmetry unless $i = j$, so

$$\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \tag{21}$$

Thus

$$cov[\boldsymbol{x}] = \mathbb{E}\boldsymbol{x}^2 - (\mathbb{E}\boldsymbol{x})^2 = \boldsymbol{\Sigma} \tag{22}$$

- *significant limitations*

  1. number of independent parameters grows quadratically with D
  2. we can assume $\boldsymbol{\Sigma}$ to be diagonal or to be proportional to the identity matrix.
  3. Only has a single maximum

  Thus the Gaussian distribution can be both too flexible in the sense of too many parameters, while also being too limited in the range of distributions that it can adequately represent.

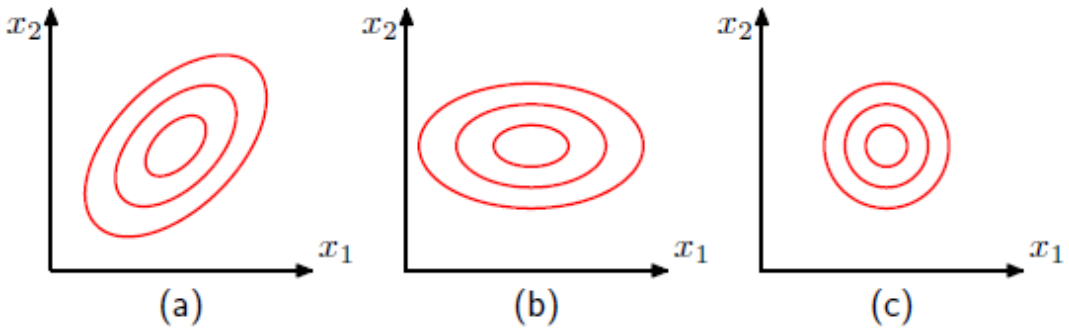

Figure 1: the covariance matrix of (a) is general form, of (b) is diagonal and of (c) is proportiional to the identity matrix

- Note that the exponent in a general Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written

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

  where we can directly get the $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$.

- *Conditional Gaussian distribution*
  If two sets of variables are joint Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian.
  Consider $\boldsymbol{x} \in \mathbb{R} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and we partition $\boldsymbol{x}$ into two subsets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, so that $\boldsymbol{x} = (\boldsymbol{x}_a^T, \boldsymbol{x}_b^T)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_a^T, \boldsymbol{\mu}_b^T)^T$ and

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

  We denote the covariance matrix by *precision matrix*

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

  where

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

  Using the equation 38 we can obtain the result of $\boldsymbol{\Lambda}_{aa}, \boldsymbol{\Lambda}_{ab}, \boldsymbol{\Lambda}_{ba}, \boldsymbol{\Lambda}_{bb}$ and $\boldsymbol{\Lambda}_{ab} = \boldsymbol{\Lambda}_{ba}, \boldsymbol{\Lambda}_{aa} = \boldsymbol{\Lambda}_{aa}^T, \boldsymbol{\Lambda}_{bb} = \boldsymbol{\Lambda}_{bb}^T$. So we obtain

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) - (\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\ &\quad - \frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \\ &= -\frac{1}{2}\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)) + const \end{aligned}$$

  Thus

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} \tag{23}$$
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \tag{24}$$

  or

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \tag{25}$$
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \tag{26}$$

  Thus

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{27}$$

  Note that $\boldsymbol{\Sigma}_{a|b}$ is independent of $\boldsymbol{x}_b$ and $\boldsymbol{\mu}_{a|b}$ is a linear function of $\boldsymbol{x}_b$. This represents an example of *linear-Gaussian* model.

- *Marginal Gaussian distribution*

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b)d\boldsymbol{x}_b = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{28}$$

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

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

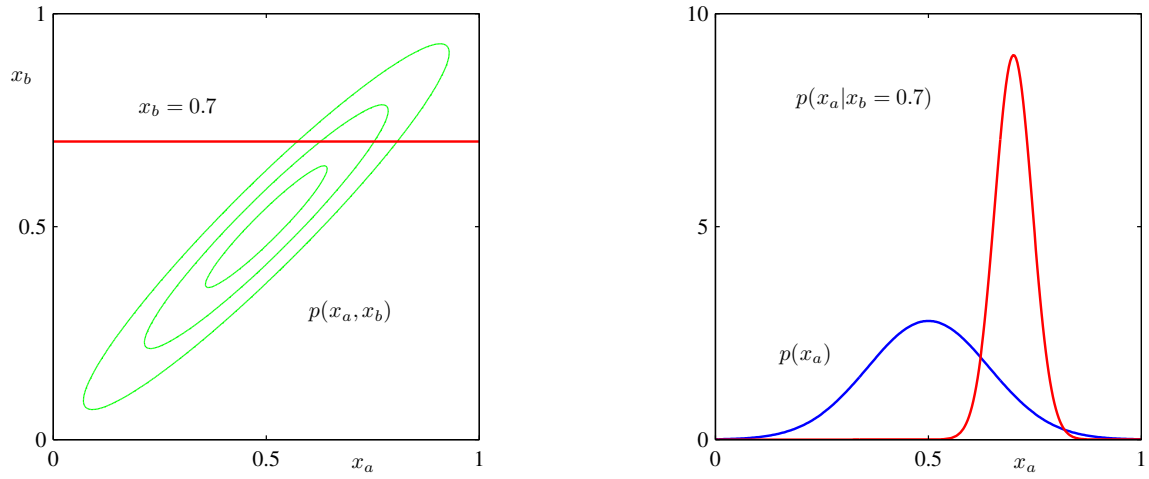

Figure 2: The conditional Gaussian distribution and the marginal Gaussian distribution

we can easily get

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

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}$$

- Given a data set $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)^T$, using the maximum likelihood function method, we can obtain

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \tag{29}$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x} - \boldsymbol{\mu}_{ML})(\boldsymbol{x} - \boldsymbol{\mu}_{ML})^T \tag{30}$$

and if we evalute the expectation of the solution, we can obtain the following results

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

So

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

- Using the equation 29, we can rewritten it as below

$$\begin{aligned}\boldsymbol{\mu}_{ML}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \\ &= \frac{1}{N}\boldsymbol{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\boldsymbol{x}_n \\ &= \frac{1}{N}\boldsymbol{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{ML}^{(N-1)} \\ &= \boldsymbol{\mu}_{ML}^{(N-1)} + \frac{1}{N}(\boldsymbol{x}_N - \boldsymbol{\mu}_{ML}^{(N-1)}) \end{aligned} \tag{31}$$

  However, we will not always be able to derive a sequential algorithm like this, which leads us to the *Robbins-Monro algorithm*, which is introduced in Appendix.
  By definition, we have

$$-\frac{\partial}{\partial\theta}\Big(\frac{1}{N}\sum_{n=1}^{N} lnp(\boldsymbol{x}_n|\theta)\Big)\Big|_{\theta_{ML}} = 0$$

  Exchanging the derivative and the summation, taking the limit $N \to \infty$, we have

$$-lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}lnp(\boldsymbol{x}_N|\theta) = \mathbb{E}_x\Big[-\frac{\partial}{\partial\theta}lnp(\boldsymbol{x}|\theta)\Big]$$

  So we can get

$$\theta_N = \theta_{N-1} - a_{N-1}\frac{\partial}{\partial\theta_{N-1}}lnp(\boldsymbol{x}_N|\theta_{N-1}) \tag{32}$$

  which equals to

$$\theta_N = \theta_{N-1} - a_{N-1}\frac{x-\mu_{ML}}{\sigma^2}$$

  If we choose $a_{N-1} = \sigma^2/N$, we can obtain the equation 31.

- We suppose the variance σ^2 is known and we consider the task of inferring the mean μ given a set of N observations $\boldsymbol{X} = (x_1, x_2, \cdots, x_N)$. The likelihood function is given by

$$p(\boldsymbol{X}|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}}e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}$$

  If we choose a prior $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$, the posterior distribution is given by

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

  where

$$\begin{aligned}\mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML} \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\end{aligned}$$

  which can be easily get by completing the square in the exponent.

- *student's t-distribution*

$$St(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\Big(\frac{\lambda}{\pi\nu}\Big)^{1/2}\Big[1+\frac{\lambda(x-\mu)^2}{\nu}\Big]^{-(\nu+1)/2} \tag{33}$$

where μ is the mean, the λ is sometimes called the precision of t-distribution, the ν is the degrees of freedom.

$$\begin{aligned} \mathbb{E}(x) &= \mu && if \quad \nu > 1 \\ var(x) &= \frac{\nu}{\nu-2}\lambda^{-1} && if \quad \nu > 2 \\ mode(x) &= \mu \end{aligned}$$

Comparing with Gaussian distribution, the student's t-distribution could be more robust for the outliers.

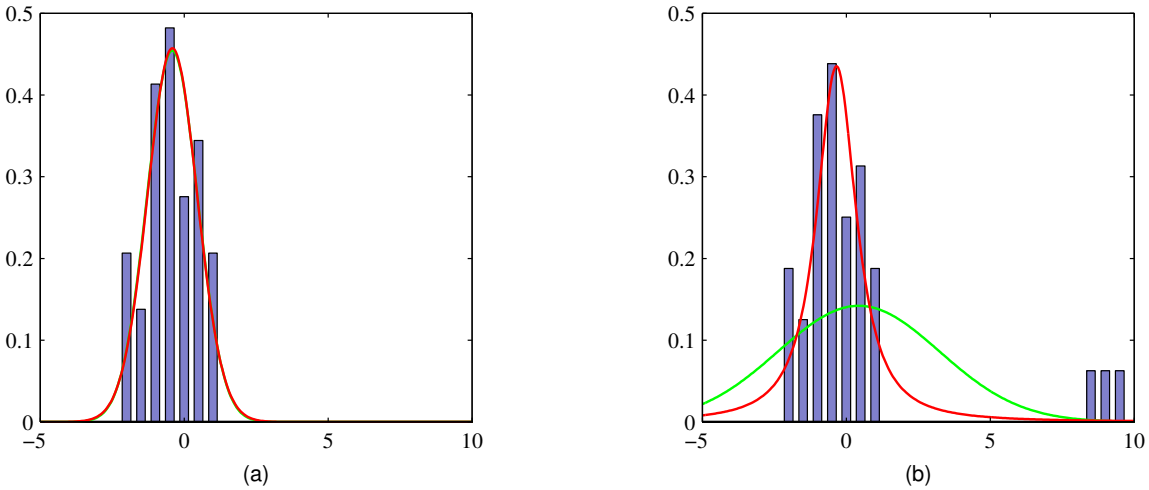

Figure 3: The robustness of t-distribution

4 The Nonparametic Methods

- *histogram methods for density estimation*

$$p_i = \frac{n_i}{N\Delta_i} \tag{34}$$

in which we often have $\Delta_i = \Delta$. Δ_i is the width of the i-th bin, n_i is the number of observations of x falling in the i-th bin and N is the total number of observations. The properties of this method is below:

1. Data can be discarded after the histogram has been computed.
2. If the data are arriving sequentially.

And the limitations are below

1. the results is dependent on the choice of edge locations and the width of bins.

2. the estimated density has discontinuities due to the bin edge.
3. Cannot be used for high dimensionality because of curse of dimensionality.

- $$p(x) = \frac{K}{NV} \tag{35}$$

  where K is the number of points lieing inside $\mathcal{R}$ and V is the volume of $\mathcal{R}$. Note that this two parameters are contradictory. We want the V can be enough small but we also want the K is large enough. If we fix the K, this method is called *K-nearest-neighbour* and if we fix the V, this method is called *kernel approach.*

5 Appendix

- *Gamma Function*
$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du \tag{36}$$
At below we prove the relation $\Gamma(x+1) = x\Gamma(x)$
$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty u^x e^{-u} du \\ &= -u^x e^{-u}\Big|_0^\infty + \int_0^\infty x u^{x-1} e^{-u} du \\ &= 0 + x\Gamma(x) \\ &= x\Gamma(x)\end{aligned}$$

- *The expression of Lagrange Multipliers in constant, vector and matrix*
For the below problem, which is expressed by exercise 2.14
$$\begin{aligned} max \quad & H[\boldsymbol{x}] = -\int p(\boldsymbol{x}) lnp(\boldsymbol{x}) d\boldsymbol{x} \\ s.t. \quad & \int p(\boldsymbol{x}) d\boldsymbol{x} = 1 \\ & \int p(\boldsymbol{x}) \boldsymbol{x} d\boldsymbol{x} = \boldsymbol{\mu} \\ & \int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T d\boldsymbol{x} = \boldsymbol{\Sigma} \end{aligned}$$
where the first constraints is constant, the second constraints is vector and the third constraints is matrix. The Lagrange function of the uppon problem is below
$$\begin{aligned} g(\lambda, \boldsymbol{m}, \boldsymbol{L}) = H[\boldsymbol{x}] &+ \lambda\Big(\int p(\boldsymbol{x}) d\boldsymbol{x} - 1\Big) + \boldsymbol{m}^T\Big(\int p(\boldsymbol{x}) \boldsymbol{x} d\boldsymbol{x} - \boldsymbol{\mu}\Big) \\ &+ tr\Big(\boldsymbol{L}\big(\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T d\boldsymbol{x} - \boldsymbol{\Sigma}\big)\Big) \end{aligned} \tag{37}$$

- *The inverse matrix of a partitioned matrix*
$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix} \tag{38}$$

where we have defined

$$\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \tag{39}$$

which is known as *Schur complement*. The process to get the equation 38 is below. We denote

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{pmatrix}$$

Thus

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I_A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_D} \end{pmatrix}$$

which equals to

$$\begin{aligned} \boldsymbol{A}\boldsymbol{E} + \boldsymbol{B}\boldsymbol{G} &= \boldsymbol{I_A} \\ \boldsymbol{C}\boldsymbol{E} + \boldsymbol{D}\boldsymbol{G} &= \boldsymbol{0} \\ \boldsymbol{A}\boldsymbol{F} + \boldsymbol{B}\boldsymbol{H} &= \boldsymbol{0} \\ \boldsymbol{C}\boldsymbol{F} + \boldsymbol{D}\boldsymbol{H} &= \boldsymbol{I_D} \end{aligned}$$

if we assume $\boldsymbol{D}$ and $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ are invertible, solve this linear equation, we can easily get the equation 38. Note that if we assume $\boldsymbol{A}$ and $\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{D}$ are invertible, the solution will be in different form. In fact, the following conditions are equivalent:

1. the original matrix is invertible.
2. $\boldsymbol{D}$ and $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ are invertible
3. $\boldsymbol{A}$ and $\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{D}$ are invertible

which can be proved by (2) $\Rightarrow$ (1), (3) $\Rightarrow$ (1), (1) $\Rightarrow$ (2) and (1) $\Rightarrow$ (3).

- *Woodbury matrix identity*

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} \tag{40}$$

where $\boldsymbol{A}$ and $\boldsymbol{C}$ are invertible square matrix and $\boldsymbol{U}$ and $\boldsymbol{V}$ can be non-square matrix.

- *Robbins-Monro algorithm*
Assume that we have a function $M(\theta)$ and a constant α, such that the equation $M(\theta) = \alpha$ has a unique root at θ^*. It is assumed that while we cannot directly observe the function $M(\theta)$, we can instead obtain measurements of the random variable $N(\theta)$ where $\mathbb{E}[N(\theta)] = M(\theta)$.

$$\theta_{n+1} = \theta_n - a_n(N(\theta_n) - \alpha) \tag{41}$$

The conditions that guarantee this algorithm convergence is below

1. $N(\theta)$ is uniformly bounded
2. $M(\theta)$ is nondecreasing
3. $M'(\theta^*)$ exists and is positive

4. the sequence a_n satisfies the following requirements

$$lim_{N\to\infty} a_N = 0$$
$$\sum_{N=1}^{\infty} a_N = \infty$$
$$\sum_{N=1}^{\infty} a_N^2 < \infty$$

The sequence of a_n which was suggested by Robbins-Monre, have the form $a_n = a/n$ for $a > 0$