Probability Theory

Pekka Marttinen

Aalto University

Contents of this lecture

In this chapter we will

- look at parametric probability density functions
- study parameter estimation from data (maximum likelihood estimation)
- apply prior knowledge to parameter estimation using the Bayes theorem
- introduce the multivariate Gaussian distribution for vector data

Basic distributions in one dimension 1/3

- When we wish to form models describing data, they are usually statistical (e.g. regression, classification, clustering, ...)
- This calls for probability theory
- Its basic concept is the probability distribution (*cumulative distribution function*) of a scalar random variable x:

$$F(x_0) = P(x \leq x_0)$$

  which gives the probability that $x \leq x_0$.
- The function $F(x_0)$ is monotonically increasing and $0 \leq F(x_0) \leq 1$.

Basic distributions in one dimension 2/3

- The *probability density function* of x when $x = x_0$:

$$p(x_0) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}\Big|_{x=x_0}$$

is the derivative of function $F(x)$ when $x = x_0$.

- Equivalently:

$$F(x_0) = \int_{-\infty}^{x_0} p(x)\mathrm{d}x$$

Basic distributions in one dimension 3/3

- Density function has the following properties:

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} x p(x) \mathrm{d}x = \mu = \mathrm{E}\{x\}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x = \mathrm{E}\{(x-\mu)^2\} = \sigma^2 = \mathrm{Var}\{x\}$$

- Here μ is the expected value (mean) and σ the standard deviation of x.
- A useful identity: $\mathrm{Var}\{x\} = \mathrm{E}\{x^2\} - \mathrm{E}\{x\}^2$

Basic distributions in one dimension: Examples

- The most common density functions are assumed to be familiar: *normal distribution* (a.k.a. Gaussian distribution)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

exponential distribution

$$p(x) = \lambda e^{-\lambda x}$$

and *uniform distribution* (in $[a, b]$)

$$p(x) = \frac{1}{b-a} \text{ when } x \in [a, b] \text{ and } 0 \text{ otherwise}$$

Normal distribution p(x)

Parameters μ (mean) and σ (standard deviation, σ^2 variance)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

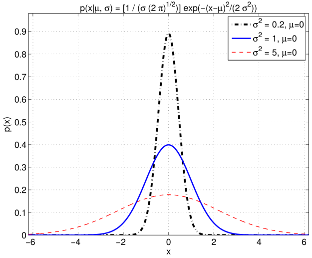

Figure: Normal distribution with 3 different values for σ.

Exponential distribution p(x)

- Parameter λ ($1/\mu$, where μ = mean)

$$p(x) = \lambda e^{-\lambda x}$$

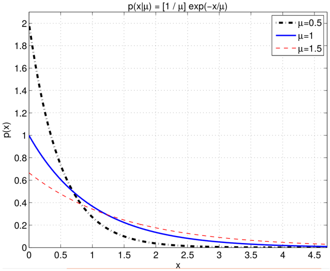

Figure: Exponential distribution with 3 different values for λ.

Uniform distribution p(x)

- Uniform in $[a, b]$

$$p(x) = \frac{1}{b-a} \text{ when } x \in [a, b] \text{ and 0 otherwise}$$

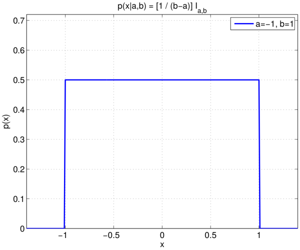

Figure: Uniform distribution.

Maximum likelihood principle: Estimation of parametric density functions 1/4

- If we have a data set/matrix $\boldsymbol{X}$ consisting of data items (vectors), how do we get their distribution?
- Knowing the distribution would be very useful, because then we'd be able to answer the following question: given a new vector $\boldsymbol{x}$, what is the probability that it belongs to the same data set (or a subset of) $\boldsymbol{X}$?
- Classification is often based on distribution estimation

Estimation of parametric density functions 2/4

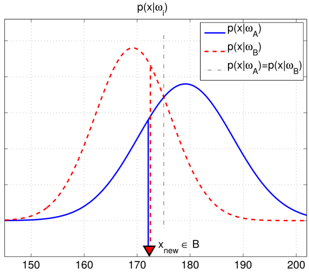

Figure: A classification example. $\boldsymbol{x}_A$ is a set of female height measurements (cm), $\boldsymbol{x}_B$ a set of male height measurements. Shown are estimated normal distributions $p_A(x) \equiv p(x|\omega_A)$ and $p_B(x) \equiv p(x|\omega_B)$. The distributions can be used to classify a new data point $p(x_{new}|\omega_B) > p(x_{new}|\omega_A)$. The observation x_{new} would be classified as a female.

Estimation of parametric density functions 3/4

- One possibility is to use a histogram to estimate the density: however, this requires much data and does not work for high-dimensional data.
- Usually a better approach is *parametric density estimation*
- We first assume that the density function $p(\boldsymbol{x})$ has a particular shape (e.g. a normal distribution) and then try to estimate its *parameters* - for a normal distribution, these would be the mean μ and the variance σ^2 (or STD σ).
- Instead of estimating the sizes of a large number of bins in a histogram, it is sufficient to learn just a few parameters to describe the distribution of data!

Histogram vs. parametric density estimation

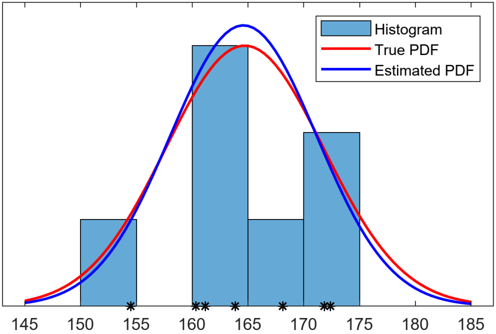

Figure: Histogram vs. parametric density estimate when the number of data points is equal to $N = 7$. The data points are shown by the asterisks.

Estimation of parametric density functions 4/4

- Let us denote the ordered set of unknown parameters with vector Θ and the density function with $p(\boldsymbol{x}|\Theta)$
- What this means: the argument of the function consist of the vector elements $x_1, \ldots, x_d$ but the function also depends on the parameters (elements of Θ)
- The general shape of the function is assumed known except for the values of the parameters Θ
- How can we learn an estimate $\hat{\Theta}$?

Maximum likelihood principle 1/3

- *Maximum likelihood* method: select the parameter vector Θ that maximizes the joint density of the data set, the so-called *likelihood function* $L(\Theta)$

$$L(\Theta) = p(\boldsymbol{X}|\Theta) = p(\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(n)|\Theta)$$

- Here $\boldsymbol{X}$ is the whole data set (matrix) and $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(n)$ are the individual data items.
- Idea: *choose the parameter values so that they give the observed data set as high probability as possible*
- Note that when we insert the numeric data $\boldsymbol{X}$ into this function, it is no longer a function of $\boldsymbol{x}$, but only of Θ

Maximum likelihood principle 2/3

Assuming the data items $(x(1), \ldots, x(n))$ are independent, the likelihood can be written as a product

$$L(\Theta) = p(\boldsymbol{X}|\Theta) = \prod_{j=1}^{n} p(\boldsymbol{x}(j)|\Theta)$$

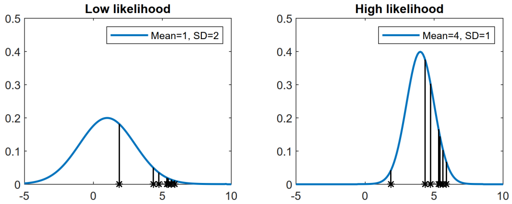

Figure: Data $(N = 7)$ are shown by asterisks on the x-axis, and assumed to follow a normal distribution. The likelihood for a given combination of parameter values is obtained by multiplying the lengths of the black lines.

Maximum likelihood principle 3/3

- Parameters are found by maximizing ("zero point of derivative")

$$\frac{\partial}{\partial\Theta} p(\boldsymbol{X}|\Theta)\bigg|_{\Theta=\hat{\theta}_{ML}} = 0$$

- Most of the time the logarithm of the likelihood function, $\ln L(\Theta)$, is used because (a) this simplifies computation (the product of density functions becomes a sum) and (b) $L(\Theta)$ and $\ln L(\Theta)$ have the same maximum.

Maximum likelihood principle: Normal distribution example 1/3

- Example: we have 1-D (scalar) data, where the data matrix only contains scalars $x(1), \ldots, x(n)$. The samples are assumed to be independent.
- Let's assume that the samples are normally distributed, but we don't know their expected value μ nor variance σ^2. We'll use the maximum likelihood method to calculate these.
- Likelihood function for "the first data point":

$$p(x(1) \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}[x(1) - \mu]^2\right]$$

Normal distribution example 2/3

- Likelihood function $L(\Theta)$ for the whole data set:

$$p(\boldsymbol{X} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2\right]$$

- Let's take the logarithm $\ln L(\Theta)$:

$$\ln p(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[x(j)-\mu]^2$$

- Let's find the maximum by setting the derivative to 0, which yields an equation that can be used to calculate μ:

$$\frac{\partial}{\partial\mu}\ln p(\boldsymbol{X} \mid \mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{j=1}^{n}[x(j)-\mu] = 0$$

Normal distribution example 3/3

- Let's solve for μ, which gives the mean of the sample

$$\mu = \hat{\mu}_{ML} = \frac{1}{n} \sum_{j=1}^{n} x(j)$$

- The corresponding equation for the variance σ^2 is

$$\frac{\partial}{\partial \sigma^2} \ln p(\boldsymbol{X} \mid \mu, \sigma^2) = 0$$

which results in an ML estimate corresponding to the sample variance

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{j=1}^{n} [x(j) - \hat{\mu}_{ML}]^2$$

Regression fitting: ML estimation for linear regression 1/4

- Example: linear regression

$$y(i) = \theta_0 + \sum_{i=1}^{d} \theta_j x_j(i) + \epsilon(i) = \theta_0 + \Theta^T \boldsymbol{x}(i) + \epsilon(i)$$

- A known way of solving for the parameters $\hat{\Theta}$ is the *least-squares method* (LSM): minimize Θ in

$$\frac{1}{n} \sum_{i=1}^{n} [y(i) - f(\boldsymbol{x}(i), \Theta)]^2$$

Example: ML estimation for linear regression 2/4

- This is in fact a *ML estimate* given the following condition: if the regression error $\epsilon(i)$ (for all i) is independent of the value of $\boldsymbol{x}(i)$ and normally distributed with expected value 0 and standard deviation σ. (A very natural assumption!)
- Then the ML estimate for parameter Θ is given by the likelihood function (here $Y = y(i)_i$)

$$p(\boldsymbol{X}, Y|\Theta) = p(\boldsymbol{X})p(Y|\boldsymbol{X}, \Theta) = p(\boldsymbol{X})\prod_{i=1}^{n} p(y(i)|\boldsymbol{X}, \Theta)$$

where we have just applied the formula for conditional probability, noting that $\boldsymbol{X}$ is independent of the regression model parameters, and assuming that the values of $y(i)$ at different measurement points are conditionally independent (standard practice in ML estimation)

ML estimation for linear regression 3/4

- We have the logarithm

$$\ln p(\boldsymbol{X}, Y|\Theta) = \ln p(\boldsymbol{X}) + \sum_{i=1}^{n} \ln p(y(i)|\boldsymbol{X}, \Theta)$$

- The distribution of $y(i)$ equals the distribution of $\epsilon(i)$ except that the mean has now moved to $f(\boldsymbol{x}(i), \Theta)$ – in other words, the normal distribution

$$p(y(i)|\boldsymbol{X}, \Theta) = \text{constant} \times \exp(-\frac{1}{2\sigma^2}[y(i) - f(\boldsymbol{x}(i), \Theta)]^2)$$

ML estimation for linear regression 4/4

- Finally, we arrive at the logarithm of the likelihood function:

$$\ln p(\boldsymbol{X}, Y|\Theta) = \ln p(\boldsymbol{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y(i) - f(\mathbf{x}(i), \Theta)]^2 + n\ln(\text{constant})$$

- Maximizing (taking the derivative with respect to Θ) makes the terms independent of Θ disappear
- This means maximizing the log-likelihood with respect to Θ is equivalent to minimizing the sum of squares! (Because $p(\boldsymbol{X})$ does not depend on Θ.)

Bayes estimation 1/6

- Bayes estimation is based on a different principle than ML estimation
- Let us assume that the unknown parameter set (vector) Θ is not constant but instead has its own distribution $p(\Theta)$
- As we get measurements/observations, the distribution of Θ grows more *exact* (e.g. its variance gets smaller)
- In practice Bayes estimation is often as simple as the ML method but gives better results
- Recall the formula for *conditional probability*:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

where A and B are events of some kind (e.g. rolls of a die)

Bayes estimation 2/6

- This can be applied to the data set and the model (parameters of the distribution) that generated the data:

$$\begin{aligned} P(model, data) &= P(model|data)P(data) = P(data, model) \\ &= P(data|model)P(model) \end{aligned}$$

- Here we have the joint distribution of the data and the model written in two different ways
- This leads us to the *Bayes formula*

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)}$$

Bayes estimation 3/6

- The idea is that we have an assumption of the probability of the model before we see any data: $P(model) \equiv prior\ probability$
- When we get some data, this gets converted into the probability $P(model|data) \equiv posterior\ probability$

Bayes estimation 4/6

- The Bayes formula tells us how to go from the prior to the posterior
- If we have a data matrix $\boldsymbol{X}$ and an unknown parameter vector Θ, the Bayes formula gives us

$$p(\Theta|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\Theta)p(\Theta)}{p(\boldsymbol{X})}$$

- NOTE: here we have used p to mark all probability distributions (density functions) and the argument tells us which p it is. This is mathematically wrong, strictly speaking, but a common shortcut.

Bayes estimation 5/6

- The Bayes formula shows us that the prior distribution is multiplied by the *likelihood function* and divided by the *probability of the data* in order to arrive at the posterior distribution
- Bayes analysis involves first "inventing" a prior distribution and then applying the Bayes formula to calculate the posterior distribution
- Often it is enough to locate the posterior maximum: maximum posterior (MAP) estimation

Bayes estimation 6/6

- This is clearly connected with ML estimation: the Bayes formula gives us

$$\ln p(\Theta|\boldsymbol{X}) = \ln p(\boldsymbol{X}|\Theta) + \ln p(\Theta) - \ln p(\boldsymbol{X})$$

  and the maximum given Θ is found with the gradient equation

$$\frac{\partial}{\partial\Theta} \ln p(\boldsymbol{X}|\Theta) + \frac{\partial}{\partial\Theta} \ln p(\Theta) = 0$$

- Compared with the ML method we now have the extra term $\partial(\ln p(\Theta))/\partial\Theta$
- This can turn out to be very useful, as we will see shortly

Example: Bayes classifier

- Suppose we know a person's height and have some *prior* probability $p(female)$ that the person is female.
- According to the Bayes rule,
  $p(female|height) \propto p(female)p(height|female)$,
  $p(male|height) \propto p(male)p(height|male)$

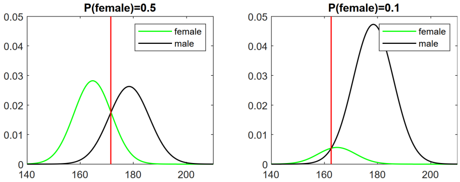

Figure: The curves show $p(height|gender)p(gender)$ for both genders and for two different priors. The red line is the decision boundary for classifying the person as male or female.

Using the Bayesian prior distribution in linear regression 1/2

- What can Bayes add to all of this?
- ML maximizes $p(\boldsymbol{X}, Y|\Theta)$. Bayes estimation maximizes $p(\Theta|\boldsymbol{X}, Y) \propto p(\boldsymbol{X}, Y|\Theta) \cdot p(\Theta)$, which after taking the logarithm equals $\ln p(\boldsymbol{X}, Y|\Theta) + \ln p(\Theta)$
- The function being maximized has a term $\ln p(\Theta)$ from the prior distribution
- The function to maximize is

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y(i) - f(\boldsymbol{x}(i), \Theta)]^2 + \ln p(\Theta)$$

- Here we can see that if the noise $\epsilon(i)$ has a very large variance σ^2, the first term is small and the estimate is strongly influenced by the prior distribution $p(\Theta)$

Using the Bayesian prior distribution in linear regression 2/2

- Whereas if σ^2 is small, the first term dominates and the prior distribution has very little effect on the result
- This seems very natural and useful!
- Example: if we have reason to assume that all parameters are normally distributed ($\mu = 0, \sigma = 1$), we have

$$p(\Theta) = \text{constant} \times \exp(-1/2 \sum_{k=1}^{K} \theta_k^2)$$

$$\ln p(\Theta) = \ln(\text{constant}) - 1/2 \sum_{k=1}^{K} \theta_k^2$$

  and maximizing the prior term results in small values for the parameters
- This usually improves prediction accuracy, especially with small data sets!

Generalization for vector data

- With vector data we have to generalize the distributions into multiple dimensions
- Let us again look at the data vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$
- Its probability distribution (cumulative distribution function) is

$$F(x_0) = P(\boldsymbol{x} \leq \boldsymbol{x}_0)$$

  where the relation "$\leq$" is understood to apply cell by cell
- The corresponding multidimensional density function $p(\boldsymbol{x})$ is its partial derivative with respect to all vector cells:

$$p(\boldsymbol{x}_0) = \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2}\cdots\frac{\partial}{\partial x_d}F(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{x}_0}$$

- In practice the density function is the more important one

Two variable normal distribution, d = 2 1/3

- A "symmetric" 2-dimensional ($d = 2$) normal distribution:

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1-\mu_1)^2+(x_2-\mu_2)^2}{2\sigma^2}}$$

where the expected value of the distribution (central point) is $\boldsymbol{m} = [\mu_1, \mu_2]$ and the standard deviation $= \sigma$ in every direction

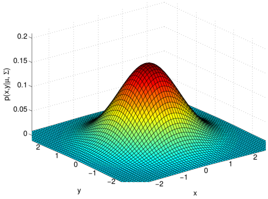

Figure: A symmetrical 2-D normal distribution.

Two variable normal distribution, d = 2 2/3

- The *2-D normal distribution* can be written in a more general form:

$$p(\boldsymbol{x}) = K \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{m}))$$

where the scaling factor K ($d = 2$) is

$$K = \frac{1}{(2\pi)^{d/2} \det(\boldsymbol{C})^{1/2}} = \frac{1}{(2\pi) \det(\boldsymbol{C})^{1/2}}$$

and $\boldsymbol{C}$ (occasionally denoted with Σ) is a covariance matrix of size (2×2) and $\boldsymbol{m} = [\mu_1 \mu_2]^T$ is the mean vector (location of peak)

Two variable normal distribution, d = 2 3/3

- Standard deviation / variance is more complicated in 2-D: variance $\sigma^2 = \mathrm{E}\{(x-\mu)^2\}$ is replaced by the covariance matrix

$$\boldsymbol{C} = \begin{bmatrix} \mathrm{E}\{(x_1-\mu_1)^2\} & \mathrm{E}\{(x_1-\mu_1)(x_2-\mu_2)\} \\ \mathrm{E}\{(x_1-\mu_1)(x_2-\mu_2)\} & \mathrm{E}\{(x_2-\mu_2)^2\} \end{bmatrix}$$

Normal distribution of d variables 1/2

- Generalization in d dimensions is straightforward: the matrix cells are $\boldsymbol{C}_{ij} = \mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = \mathrm{Cov}(x_i, x_j)$
- The covariance matrix $\boldsymbol{C}$ is a symmetrical square matrix of size $(d \times d)$
- The density function of a d-dimensional normal distribution can be written in a vector-matrix format:

$$p(\boldsymbol{x}) = K \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{m}))$$

where $\boldsymbol{m} = [\mu_1, \ldots, \mu_d]^T$ is the mean vector (central point of distribution, peak) and K is the normalizing term

$$K = \frac{1}{(2\pi)^{d/2} \det(\boldsymbol{C})^{1/2}}$$

Normal distribution of d variables 2/2

- The normalizing term K is only required to make the integral of $p(\boldsymbol{x})$ over the d-dimensional space equal to 1
- Using integration we can derive

$$\mathrm{E}\{\boldsymbol{x}\} = \boldsymbol{m} = \int_{R^d} \boldsymbol{x} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$\mathrm{E}\{(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^T\} = \boldsymbol{C}$$

- Note: even though $d > 1$, $p(\boldsymbol{x}) \in \mathbb{R}^1$: the density function is scalar valued, as matrix multiplication in the exponential function $(1 \times \underline{d})(\underline{d} \times \underline{\underline{d}})(\underline{\underline{d}} \times 1) = (1 \times 1)$

Uncorrelatedness and independence 1/3

- The cells of vector $\boldsymbol{x}$ are *uncorrelated* if their covariances are zero:

$$\mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = 0$$

- In other words, the covariance matrix $\boldsymbol{C}$ is diagonal:

$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & & \dots & \sigma_d^2 \end{bmatrix}$$

- The cells of vector $\boldsymbol{x}$ are *independent* if their *joint distribution* can be stated as the product of their *marginal distributions*:

$$p(\boldsymbol{x}) = p_1(x_1)p_2(x_2)\dots p_d(x_d)$$

Uncorrelatedness and independence 2/3

- Independence implies uncorrelatedness, but not necessarily vice versa
- If the cells of a *normally distributed vector* are uncorrelated, they are also independent
- If we have $\mathrm{E}\{(x_i - \mu_i)(x_j - \mu_j)\} = 0$ (uncorrelatedness), $\boldsymbol{C}$ is a diagonal matrix and

$$(\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{m}) = \sum_{i=1}^{d} (\boldsymbol{C}_{ii})^{-1} (x_i - \mu_i)^2$$

Uncorrelatedness and independence 3/3

- Its exponential function is

$$
\begin{aligned}
p(\boldsymbol{x}) &= K \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right) \\
&= K \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{d}(\boldsymbol{C}_{ii})^{-1}(x_i-\mu_i)^2\right) \\
&= K \cdot \exp(-\frac{1}{2}\boldsymbol{C}_{11}^{-1}(x_1-\mu_1)^2) \ldots \exp(-\frac{1}{2}\boldsymbol{C}_{dd}^{-1}(x_d-\mu_d)^2)
\end{aligned}
$$

- $p(\boldsymbol{x})$ can be expressed as the product of marginal distributions $\rightarrow$ the cells x_i are independent

Summary

In this chapter we

- looked at density functions
- studied the maximum likelihood method for parameter estimation
- applied prior knowledge (a priori) to parameter estimation using the Bayes theorem
- introduced the multivariate Gaussian distribution for vector data