Introduction to Bayesian Methods- 2

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Connection with frequentist statistics:

Consider the following identity involving the (joint and conditional) distributions of parameters and data

$$\mathbb{E}_\theta[\theta] = \int \theta \; p(\theta) d\theta = \int \int \theta \; p(\theta, \mathcal{D}) d\theta d\mathcal{D} = \int \left[\int \theta \; p(\theta|\mathcal{D}) d\theta \right] p(\mathcal{D}) d\mathcal{D}$$

$$= \mathbb{E}_\mathcal{D} \left[\mathbb{E}_\theta[\theta|\mathcal{D}] \right]$$

Prior mean

FINAL RESULT:
Posterior mean, averaged over the distribution of data

Marginalization over the distribution of data

Splitting the joint pdf

Connection with frequentist statistics - 2:

A similar identity holds when we consider the variance

$$\begin{aligned}
\mathrm{var}_\theta(\theta) &= \int \theta^2 \; p(\theta) d\theta - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \int \left[\int \theta^2 \; p(\theta|\mathcal{D}) d\theta \right] p(\mathcal{D}) d\mathcal{D} - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \mathbb{E}_\mathcal{D}\left[\mathbb{E}_\theta[\theta^2|\mathcal{D}]\right] - \left(\mathbb{E}_\theta[\theta]\right)^2 \\
&= \mathbb{E}_\mathcal{D}\left[\mathbb{E}_\theta[\theta^2|\mathcal{D}]\right] - \mathbb{E}_\mathcal{D}\left[\left(\mathbb{E}_\theta[\theta|\mathcal{D}]\right)^2\right] + \mathbb{E}_\mathcal{D}\left[\left(\mathbb{E}_\theta[\theta|\mathcal{D}]\right)^2\right] - \left(\mathbb{E}_\mathcal{D}\left[\mathbb{E}_\theta[\theta|\mathcal{D}]\right]\right)^2 \\
&= \mathbb{E}_\mathcal{D}\left[\mathrm{var}_\theta(\theta|\mathcal{D})\right] + \mathrm{var}_\mathcal{D}\left(\mathbb{E}_\theta[\theta|\mathcal{D}]\right)
\end{aligned}$$

Prior variance

posterior variance, averaged over the distribution of data

variance over data of the posterior mean

$$\mathrm{var}_\theta(\theta) \geq \mathbb{E}_\mathcal{D}\left[\mathrm{var}_\theta(\theta|\mathcal{D})\right]$$

the prior variance is greater or equal than the (mean) posterior variance !!!

The Bernstein-Von Mises theorem

- The theorem that grants convergence under very weak hypotheses is the Bernstein-Von Mises theorem. The theorem states that a posterior distribution converges in the limit of infinite data to a multivariate normal distribution centered at the maximum likelihood estimator with covariance matrix given by the normalized Fisher matrix.

- Convergence can only be defined with respect to a frequentist approach (this requires repeated, independent tests of the experimental procedure).

- In the case of nonparametric statistics and for certain probability spaces, the Bernstein-von Mises theorem usually fails.

Predictive power of the posterior distribution in the problem of coin tosses

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1,\theta|\mathcal{D})d\theta = \int_0^1 p(x=1|\theta,\mathcal{D})\; p(\theta|\mathcal{D})d\theta =$$

$$= \int_0^1 \theta\; p(\theta|\mathcal{D})d\theta = \mathbb{E}[\theta|\mathcal{D}]$$

$$= \frac{n+m}{N+m+l}$$

updated at every step; recall that

n: number of observed heads
N: total number of coin tosses
m: parameter specifying the initial prior

Generalization to more than two discrete values:
the 1-of-K coding scheme and the Dirichlet pdf

Coin tosses and Bernoulli variates

there are 2 cases that can take place, and they are mutually exclusive

- Bernoulli variates:

$$x \sim \begin{cases} P(x=1) = \theta \\ P(x=0) = 1-\theta \end{cases} \quad \Rightarrow \quad \begin{aligned} \mathbb{E}(x) &= \theta \\ \mathrm{var}(x) &= \theta \end{aligned}$$

- 1-of-2 coding scheme for Bernoulli variates:

lists of possible outcomes

$$x \quad \Rightarrow \quad (1,0)^T \text{ OR } (0,1)^T$$

- The Binomial variate as a sum of Bernoulli variates:

# of 1s, # of 0s

$$x_{\mathrm{Bin}} = \sum_{k=1}^{N} x_k \quad \Rightarrow \quad P(x_{\mathrm{Bin}}) = \frac{N!}{m!l!} x^m (1-x)^l$$

What is 1-of-K coding used for?

This kind of coding is used in Machine Learning (ML) to convert categorical information into numerical data.

This conversion is utilized in clustering method like K-means.

The case of multiple values

- 1-of-K coding scheme:

$$\mathbf{x} = (x_1, \ldots, x_k, \ldots, x_K)^T = (0, \ldots, 1, \ldots, 0)^T$$

- Constraint:

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

- Probability of each scalar component:

$$P(x_k = 1) = \theta_k; \quad \sum_{k=1}^{K} \theta_k = 1$$

- Probability of a given vector:

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}$$

The case of multiple values – 2

- Dataset of N independent observations:

$$\mathbf{x}_n = (x_{n1}, \ldots, x_{nk}, \ldots, x_{nK})^T$$

- Constraint:

$$\sum_{k=1}^{K} x_{nk} = 1$$

- Probability of each scalar component:

$$P(x_{nk} = 1) = \theta_k; \quad \sum_{k=1}^{K} \theta_k = 1 \quad \Rightarrow \quad \mathbb{E}(\mathbf{x}|\boldsymbol{\theta}) = \boldsymbol{\theta}$$

- Probability of a given vector:

$$P(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_{nk}}$$

The case of multiple values – 3

Likelihood for a dataset of size N

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

where

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

which is the number of observations of the k-th element, **all of them i.i.d.**

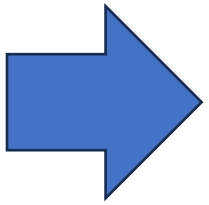
To maximize the log-likelihood we must use a Lagrange multiplier

$$\frac{\partial}{\partial\theta_j}\left[\sum_{k=1}^{K} m_k \ln\theta_k - \lambda\left(\sum_k \theta_k - 1\right)\right] = \frac{m_j}{\theta_j} - \lambda = 0 \quad \Rightarrow \quad \theta_j = \frac{m_j}{\lambda}$$

The case of multiple values – 4

From the normalization condition

$$1 = \sum_{j=1}^{K} \theta_j = \sum_{j=1}^{K} \frac{m_j}{\lambda} = \frac{N}{\lambda} \quad \Rightarrow \quad \lambda = N$$



$$\hat{\theta}_j = \frac{m_j}{N}$$

We do not need ALL the data to determine the parameters, we only need the m_j's, which are thus an example of **sufficient statistic**

The case of multiple values – 5

Except for the normalization, the likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{m_k}$$

is proportional to the multinomial distribution of the quantities m_j's

$$\text{Mult}(m_1, \ldots, m_K|\boldsymbol{\theta}) = \frac{N!}{\prod_{k=1}^{K} m_k!} \prod_{k=1}^{K} \theta_k^{m_k}$$

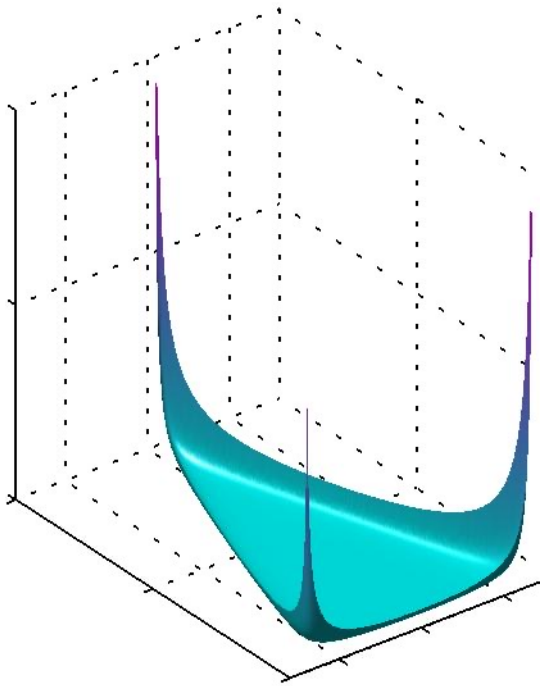
and the corresponding conjugate distribution is the Dirichlet distribution (a generalization of the Beta distribution)

$$\text{Dir}(\boldsymbol{\theta}|\mathbf{m}) = \frac{\Gamma(m_0+1)}{\prod_{k=1}^{K} \Gamma(m_k+1)} \prod_{k=1}^{K} \theta_k^{m_k} \quad \text{with } m_0 = \sum_k m_k$$

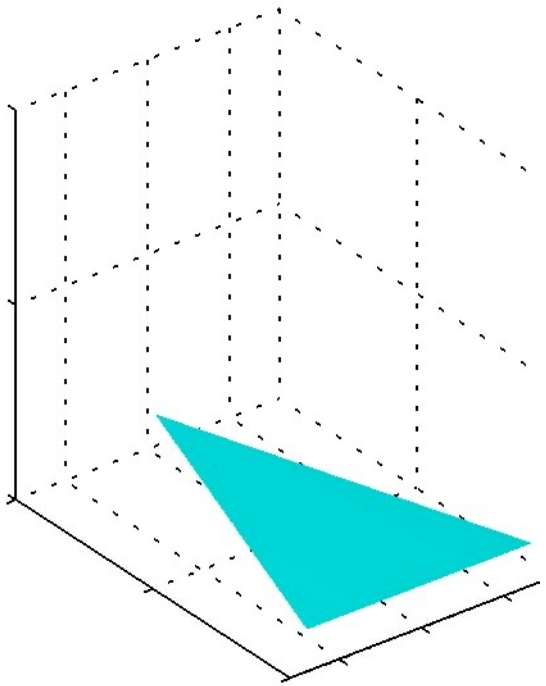
Thus, a prior Dirichlet distribution is conjugate to the posterior and we can use it just as the Beta distribution of the previous example.

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0 + 1)}{\prod_{k=1}^{K} \Gamma(\alpha_k + 1)} \prod_{k=1}^{K} \theta_k^{\alpha_k} \quad \text{with } \alpha_0 = \sum_k \alpha_k$$

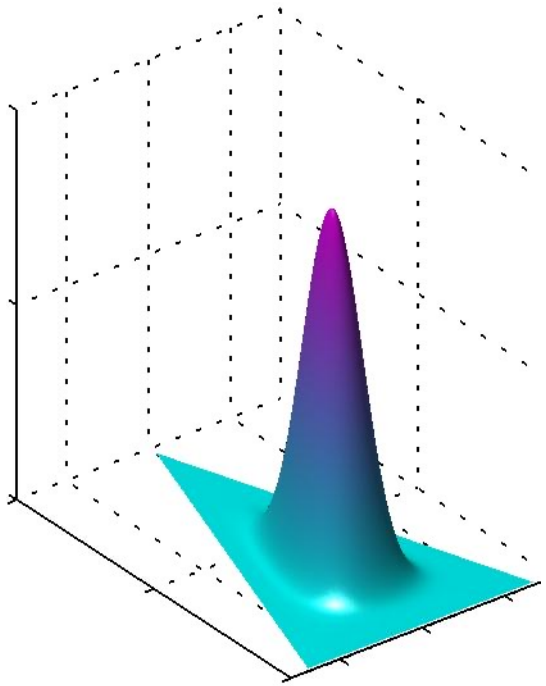
Dirichlet prior



$\alpha_k = 0.1$



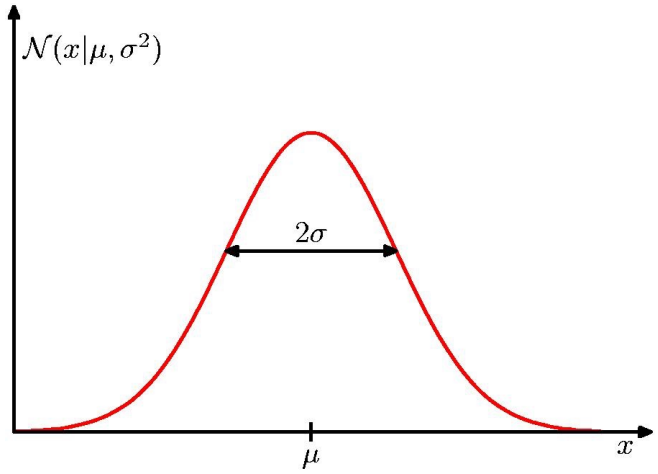
$\alpha_k = 1$



$\alpha_k = 10$

The (multivariate) Gaussian distribution:

1-dimensional Gaussian distribution



Mean

Variance

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

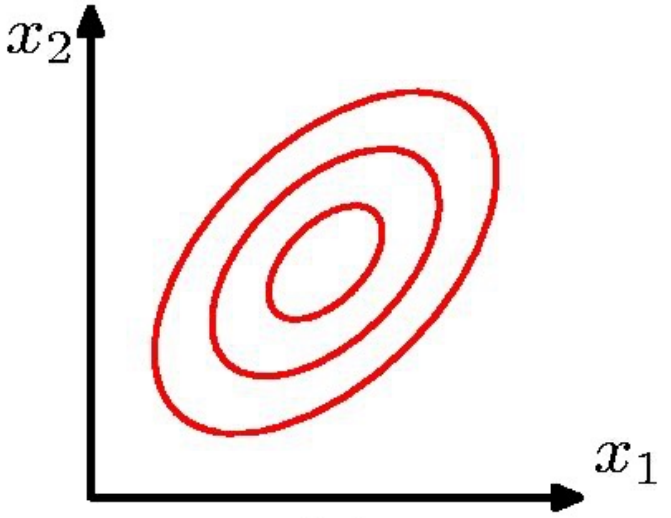
D-dimensional Gaussian distribution

Vector of means

Covariance matrix

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

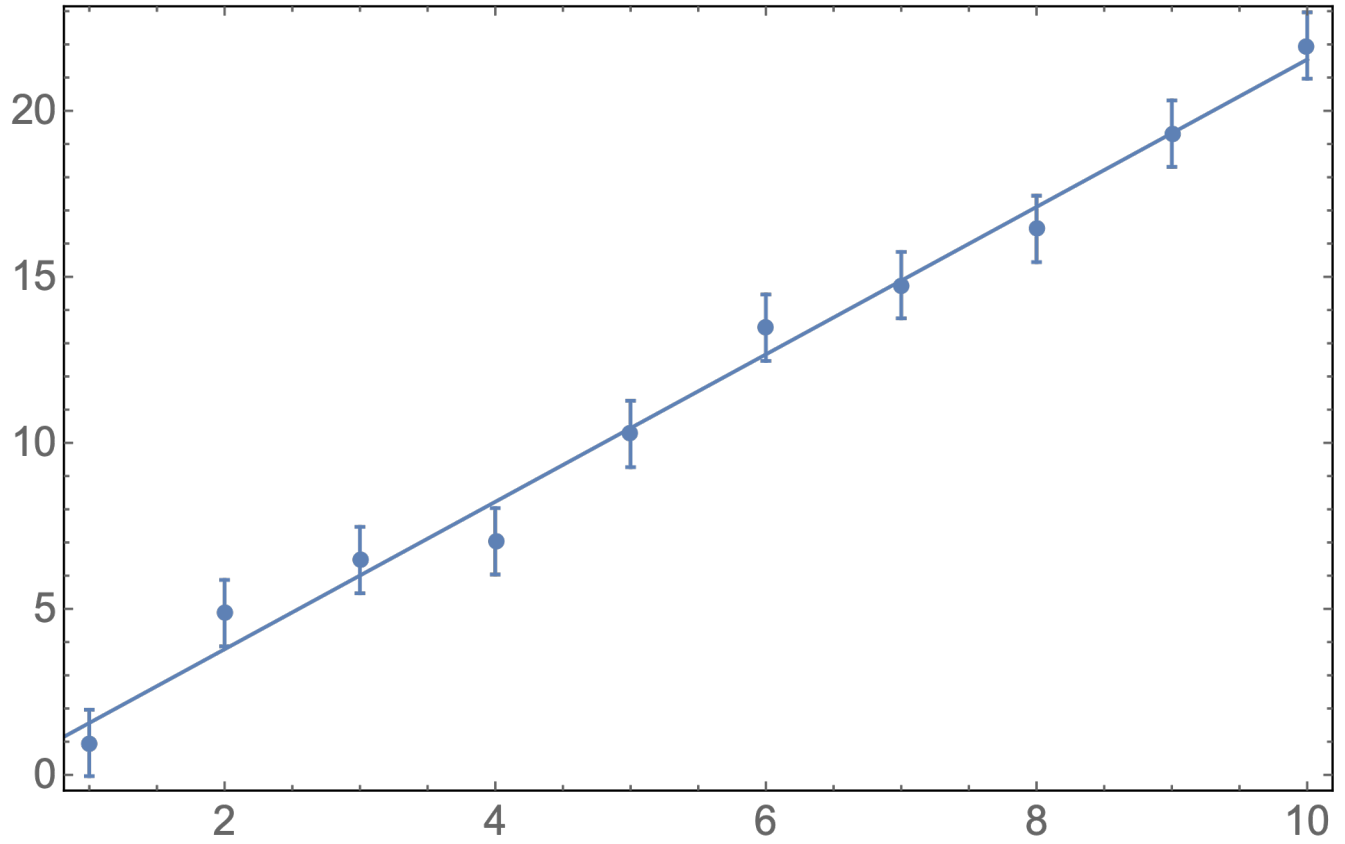
Square root of the determinant of the covariance matrix

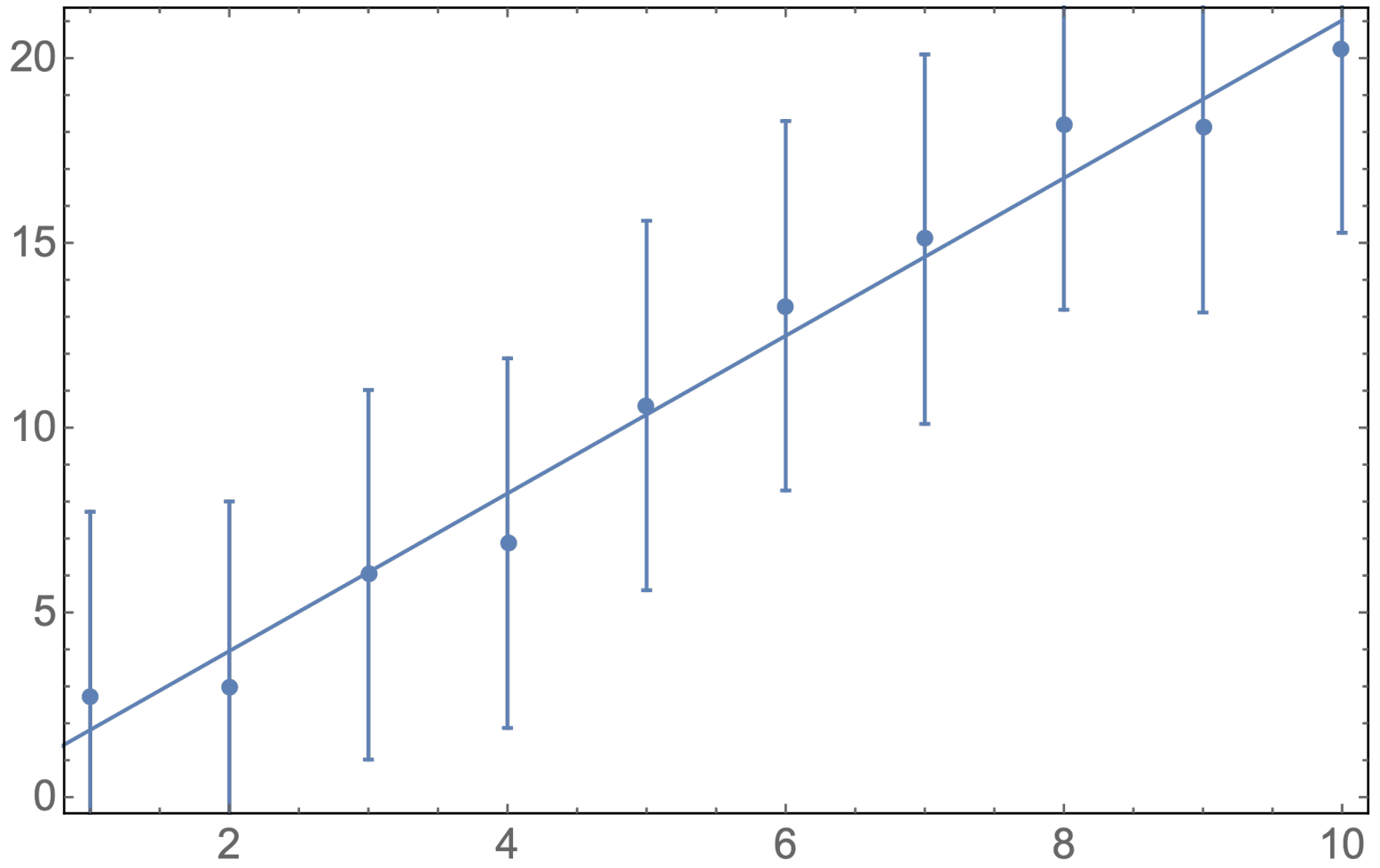


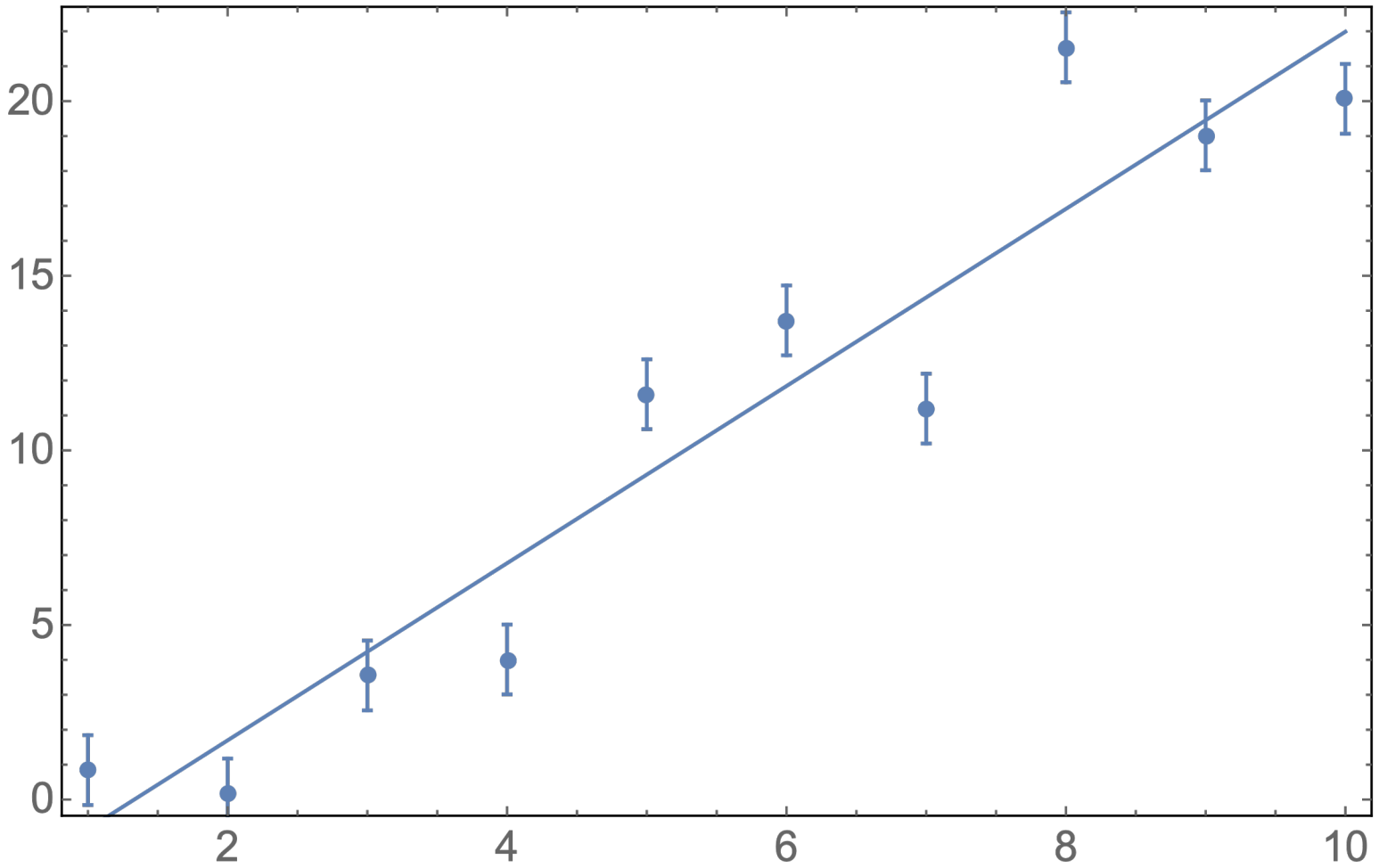
Example of Bayesian estimate using objective priors: uncalibrated Gaussian measurement uncertainties

Here, we consider the case where we must find the mean value with given measurement uncertainties that are systematically multiplied by an unknown scale factor, under the assumption of Gaussianity.

In this example we "complete the square" with "old style" methods, to provide a comparison with the methods developed later.




20
15
10
5
0
2
4
6
8
10


20
15
10
5
0
2
4
6
8
10

The likelihood has a Gaussian structure

$$P(\mathbf{d} \mid \mu, \boldsymbol{\sigma}, \alpha) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\alpha^2\sigma_k^2}} \exp\left[-\frac{(d_k - \mu)^2}{2\alpha^2\sigma_k^2} \right]$$

$$= \frac{1}{(2\pi)^{N/2}\alpha^N} \left(\prod_{k=1}^{N} \frac{1}{\sigma_k} \right) \exp\left[-\frac{1}{2\alpha^2} \sum_{k=1}^{N} \frac{(d_k - \mu)^2}{\sigma_k^2} \right]$$

we must rearrange the exponent as usual ...

$$\sum_{k=1}^{N} \frac{(d_k - \mu)^2)}{\sigma_k^2} = \sum_{k=1}^{N} \frac{d_k^2}{\sigma_k^2} - 2\mu \sum_{k=1}^{N} \frac{d_k}{\sigma_k^2} + \mu^2 \sum_{k=1}^{N} \frac{1}{\sigma_k^2} = \frac{ND}{\sigma_M^2} - 2\mu \frac{NM}{\sigma_M^2} + \mu^2 \frac{1}{\sigma_M^2}$$

$$= \frac{N}{\sigma_M^2} \left(D - 2\mu M + \mu^2 \right)$$

$$\text{where } \frac{1}{\sigma_M^2} = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sigma_k^2}; \quad M = \sum_{k=1}^{N} \frac{d_k}{\sigma_k^2} \Big/ \sum_{k=1}^{N} \frac{1}{\sigma_k^2}; \quad D = \sum_{k=1}^{N} \frac{d_k^2}{\sigma_k^2} \Big/ \sum_{k=1}^{N} \frac{1}{\sigma_k^2}$$

therefore, the likelihood is

$$P(\mathbf{d}|\mu, \sigma, \alpha) = \frac{1}{(2\pi)^{N/2} \alpha^N} \left(\prod_{k=1}^{N} \frac{1}{\sigma_k} \right) \exp \left[-\frac{N}{2\alpha^2 \sigma_M^2} (D - 2\mu M + \mu^2) \right]$$

Now we estimate the scale factor from Bayes' theorem

$$p(\alpha|\mathbf{d},\boldsymbol{\sigma}) = \frac{p(\mathbf{d}|\alpha,\boldsymbol{\sigma})}{\int_\alpha p(\mathbf{d}|\alpha',\boldsymbol{\sigma})p(\alpha')d\alpha'}p(\alpha)$$

however, we need first to marginalize the likelihood with respect to the mean, which in this case is a ***nuisance parameter***

we take a uniform prior for the mean (a Jeffrey's prior, see later)

$$P(\mathbf{d}\,|\,\boldsymbol{\sigma},\alpha) = \int_\mu P(\mathbf{d}\,|\,\mu,\boldsymbol{\sigma},\alpha)P(\mu|\boldsymbol{\sigma},\alpha)d\mu$$

$$= \frac{1}{W}\int_{\mu_{\min}}^{\mu_{\max}} P(\mathbf{d}\,|\,\mu,\boldsymbol{\sigma},\alpha)d\mu$$

$$\approx \frac{1}{W}\frac{1}{(2\pi)^{N/2}\alpha^N}\left(\prod_{k=1}^{N}\frac{1}{\sigma_k}\right)\int_{-\infty}^{+\infty}\exp\left[-\frac{N}{2\alpha^2\sigma_M^2}\left(D-2\mu M+\mu^2\right)\right]d\mu$$

$$W = \mu_{\max} - \mu_{\min}$$

as usual ...

$$D - 2\mu M + \mu^2 = \mu^2 - 2\mu M + M^2 + D - M^2 = (\mu - M)^2 + D - M^2$$

... therefore, the marginalized likelihood is:

$$P(\mathbf{d} \mid \boldsymbol{\sigma}, \alpha) \approx \frac{1}{W} \frac{1}{(2\pi)^{N/2} \alpha^N} \left(\prod_{k=1}^{N} \frac{1}{\sigma_k} \right) \int_{-\infty}^{+\infty} \exp\left\{ -\frac{N}{2\alpha^2 \sigma_M^2} \left[(\mu - M)^2 + D - M^2 \right] \right\} d\mu$$

$$= \frac{1}{W} \frac{1}{(2\pi)^{N/2} \alpha^N} \left(\prod_{k=1}^{N} \frac{1}{\sigma_k} \right) \exp\left(-\frac{N\left(D - M^2\right)}{2\alpha^2 \sigma_M^2} \right) \sqrt{\frac{2\pi\alpha^2 \sigma_M^2}{N}}$$

$$p(\alpha|\mathbf{d},\boldsymbol{\sigma}) = \frac{p(\mathbf{d}|\alpha,\boldsymbol{\sigma})}{\int_\alpha p(\mathbf{d}|\alpha',\boldsymbol{\sigma})p(\alpha')d\alpha'}p(\alpha)$$

$$= \frac{\frac{1}{\alpha^{N-1}}\exp\left(-\frac{N(D-M^2)}{2\alpha^2\sigma_M^2}\right)}{\int_\alpha \frac{1}{\alpha'^{N-1}}\exp\left(-\frac{N(D-M^2)}{2\alpha'^2\sigma_M^2}\right)p(\alpha')d\alpha'}p(\alpha)$$

$$P(\alpha) \propto \frac{1}{\alpha}$$

for the standard deviation we take again a Jeffreys' prior

$$p(\alpha|\mathbf{d},\boldsymbol{\sigma}) = \frac{\frac{1}{\alpha^{N-1}}\exp\left(-\frac{N(D-M^2)}{2\alpha^2\sigma_M^2}\right)\frac{1}{\alpha}}{\int_\alpha \frac{1}{\alpha'^{N-1}}\exp\left(-\frac{N(D-M^2)}{2\alpha'^2\sigma_M^2}\right)\frac{1}{\alpha'}d\alpha'}; \quad A^2 = \frac{N(D-M^2)}{2\sigma_M^2}$$

$$\Rightarrow \quad p(\alpha|\mathbf{d},\boldsymbol{\sigma}) \rightarrow \frac{\frac{1}{\alpha^N}\exp\left(-\frac{A^2}{\alpha^2}\right)}{\int_0^\infty \frac{1}{\alpha'^N}\exp\left(-\frac{A^2}{\alpha'^2}\right)d\alpha'}$$

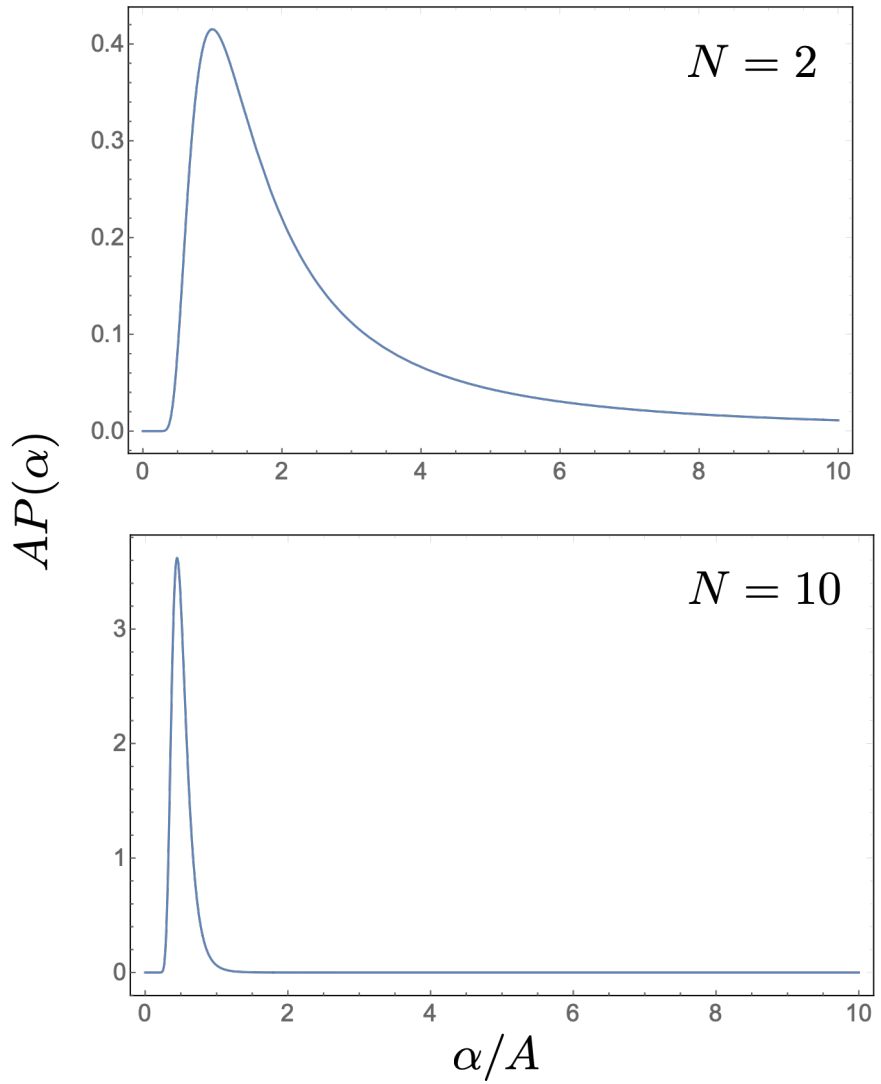
evaluation of $\int_0^\infty \frac{1}{\alpha'^N} \exp\left(-\frac{A^2}{\alpha'^2}\right) d\alpha'$

$$\frac{A^2}{\alpha^2} = x; \quad \alpha = \frac{A}{\sqrt{x}}; \quad d\alpha = -\frac{A}{2x^{3/2}} dx$$

$$\int_0^\infty \frac{x^{N/2}}{A^N} \exp(-x) \frac{A}{2x^{3/2}} dx = \frac{1}{2A^{N-1}} \int_0^\infty x^{\frac{N-1}{2}-1} \exp(-x) dx = \frac{1}{2A^{N-1}} \Gamma\left(\frac{N-1}{2}\right)$$

$$p(\alpha|\mathbf{d}, \boldsymbol{\sigma}) \to \frac{\frac{2A^{N-1}}{\alpha^N} \exp\left(-\frac{A^2}{\alpha^2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}$$

$$P(\alpha|\mathbf{d},\boldsymbol{\sigma}) = \frac{(2A^{N-1}/\alpha^N)\exp(-A^2/\alpha^2)}{\Gamma\left[(N-1)/2\right]}$$



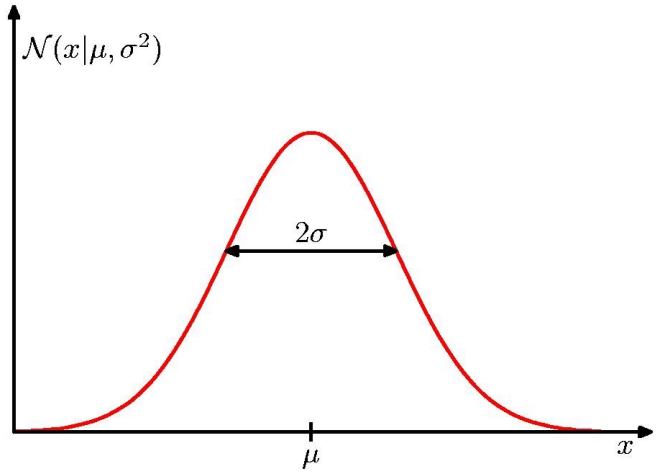
we take the MAP estimate of the scale parameter from the pdf

$$p(\alpha|\mathbf{d},\boldsymbol{\sigma}) = \frac{\dfrac{2A^{N-1}}{\alpha^N}\exp\left(-\dfrac{A^2}{\alpha^2}\right)}{\Gamma\left(\dfrac{N-1}{2}\right)}$$

$$\frac{d}{d\alpha}P(\alpha\,|\,\mathbf{d},\boldsymbol{\sigma}) \propto -\frac{N}{\alpha^{N+1}}\exp\left(-\frac{A^2}{\alpha^2}\right)+\frac{2A^2}{\alpha^{N+3}}\exp\left(-\frac{A^2}{\alpha^2}\right)=0$$

$$\Rightarrow \quad N\alpha^2 = 2A^2 \quad \Rightarrow \quad \alpha_{\mathrm{MAP}} = \sqrt{\frac{2}{N}}A = \sqrt{\frac{D-M^2}{\sigma_M^2}}$$

1-dimensional Gaussian distribution



Mean

Variance

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

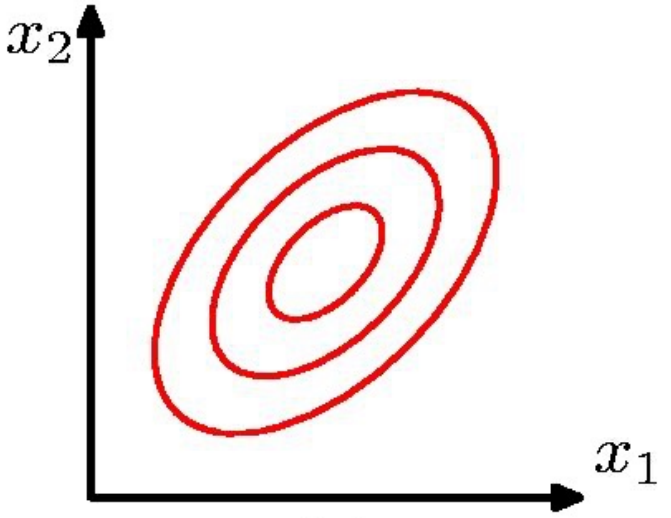
D-dimensional Gaussian distribution

Vector of means

Covariance matrix

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

Square root of the determinant of the covariance matrix



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

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

Mahalanobis distance

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

Eingenvalue equation for the **covariance matrix**

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

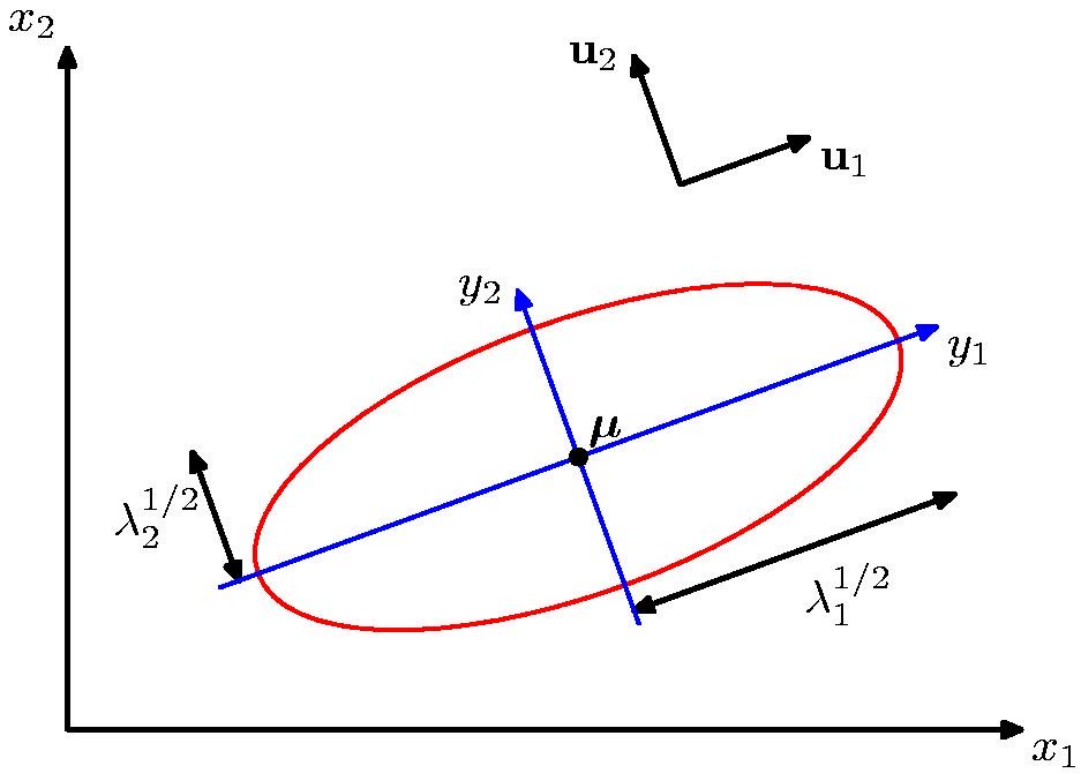
Diagonal form of the **precision matrix**

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

Mahalanobis distance in the rotated/translated reference system

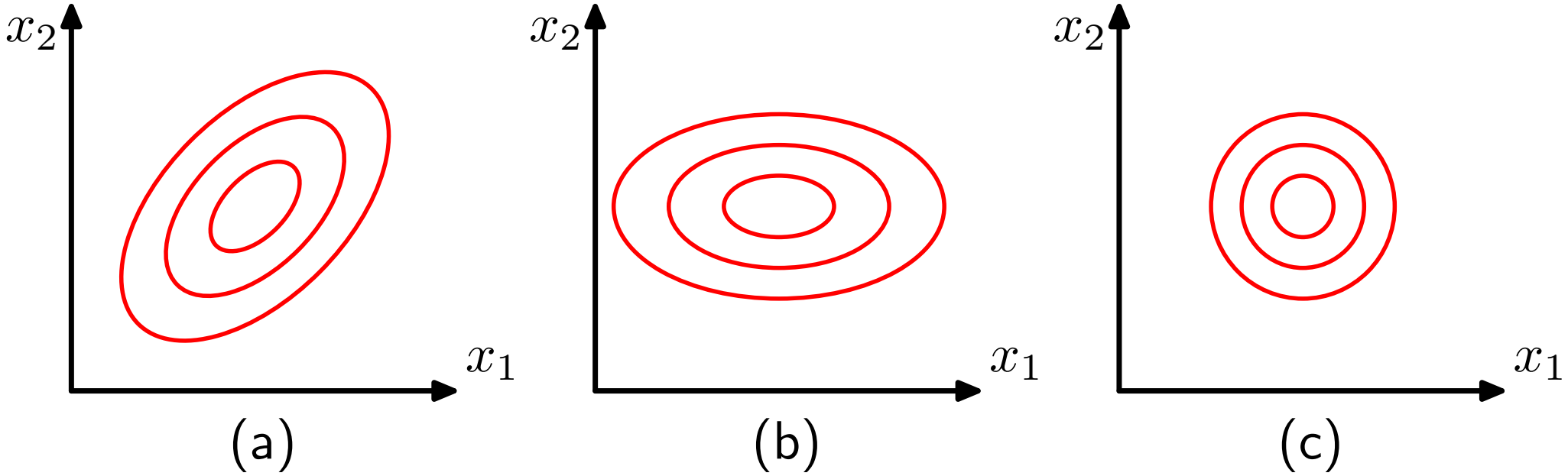
$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$

Rotated/translated coordinates



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

pdf in the rotated/translated system



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