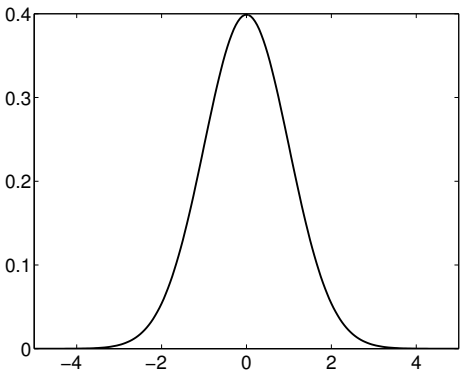
REVIEW: GAUSSIAN DISTRIBUTIONS

Gaussian density in one dimension

$$p(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$$



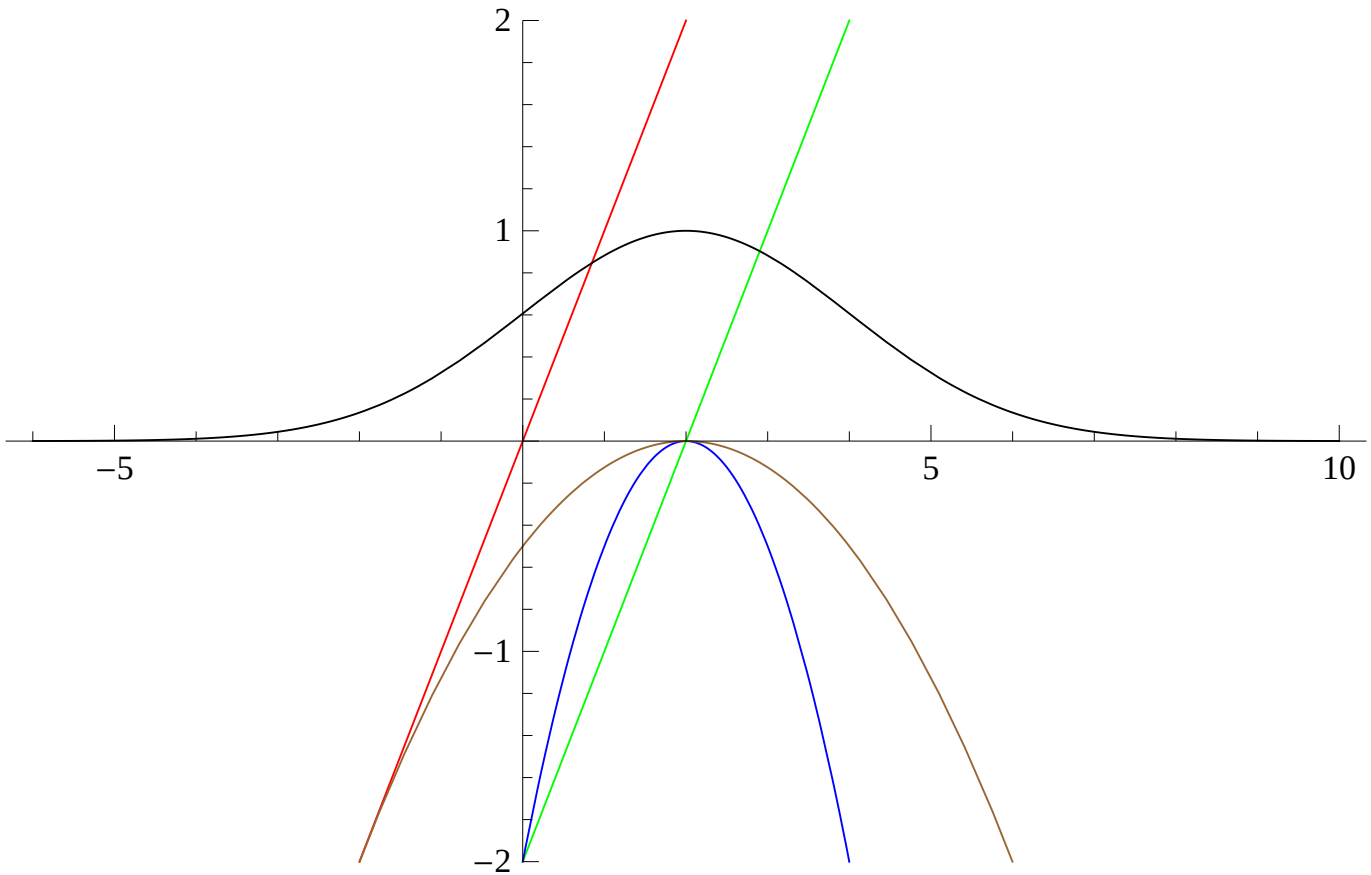
- μ = expected value of x, σ^2 = variance, σ = standard deviation
- The quotient $\frac{x-\mu}{\sigma}$ measures deviation of x from its expected value units of σ (i.e. σ defines the length scale)

Recall: Standard deviation around the mean

- Recall that the interval $[\mu-\sigma, \mu+\sigma]$ ("one standard deviation") always contains the same amount of probability mass (ca. 68.27%), regardless of the choice of μ and σ.
- Similarly, the intervall $[\mu-2\sigma, \mu+2\sigma]$ contains $\sim 95.45\%$ of the mass, and $[\mu-3\sigma, \mu+3\sigma]$ contains $\sim 99.73\%$.

Components of a 1D Gaussian

$\mu = 2, \sigma = 2$



- Red: $x \mapsto x$
- Green: $x \mapsto x - \mu$
- Blue: $x \mapsto -\frac{1}{2}(x - \mu)^2$
- Brown: $x \mapsto -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$
- Black: $x \mapsto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

Recall: Covariance

The covariance of two random variables X_1, X_2 is

$$\mathrm{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] \ .$$

If $X_1 = X_2$, the covariance is the variance: $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$.

Covariance matrix

If $X = (X_1, \ldots, X_m)$ is a random vector with values in $\mathbb{R}^m$, the matrix of all covariances

$$\mathrm{Cov}[X] := (\mathrm{Cov}[X_i, X_j])_{i,j} = \begin{pmatrix} \mathrm{Cov}[X_1, X_1] & \cdots & \mathrm{Cov}[X_1, X_m] \\ \vdots & & \vdots \\ \mathrm{Cov}[X_m, X_1] & \cdots & \mathrm{Cov}[X_m, X_m] \end{pmatrix}$$

is called the **covariance matrix** of X.

Notation

It is customary to denote the covariance matrix $\mathrm{Cov}[X]$ by Σ.

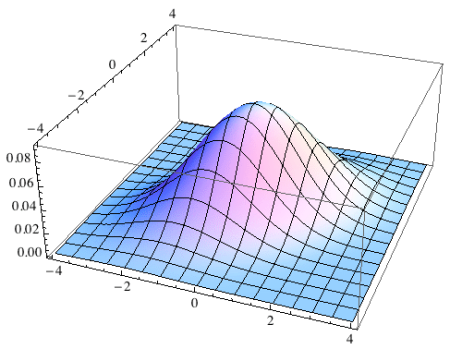
Gaussian density in m dimensions

The quadratic function

$$-\frac{(x-\mu)^2}{2\sigma^2} \quad = \quad -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

is replaced by a quadratic form:

$$p(\mathbf{x};\boldsymbol{\mu},\Sigma) := \frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\Big(-\frac{1}{2}\Big\langle(\mathbf{x}-\boldsymbol{\mu}),\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\rangle\Big)$$

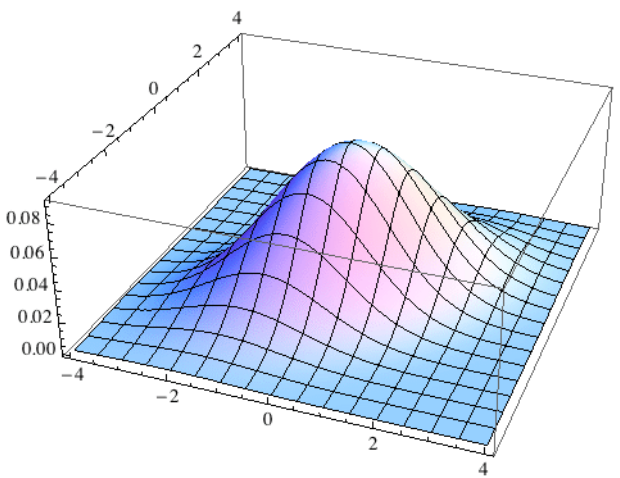


Covariance matrix of a Gaussian

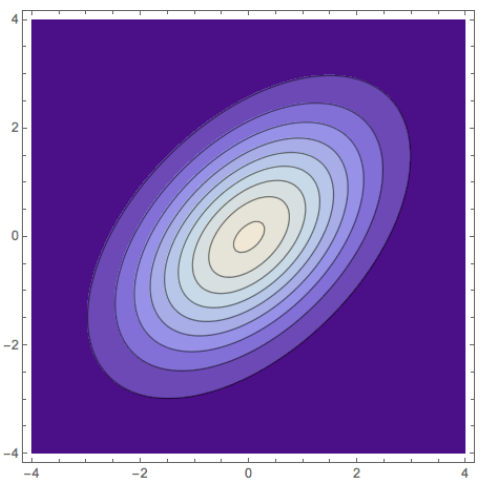
If a random vector $X \in \mathbb{R}^m$ has Gaussian distribution with density $p(\mathbf{x};\mu,\Sigma)$, its covariance matrix is $\mathrm{Cov}[X] = \Sigma$. In other words, a Gaussian is parameterized by its covariance.

GAUSSIAN DENSITY: EXAMPLE

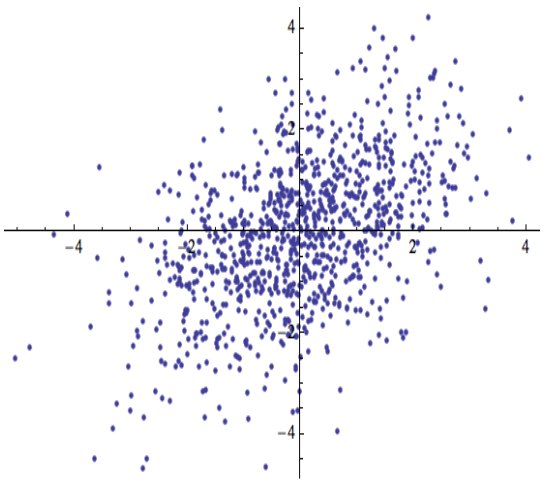
$$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \quad \text{with} \quad \boldsymbol{\mu} = (0, 0) \quad \text{with} \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$



Density

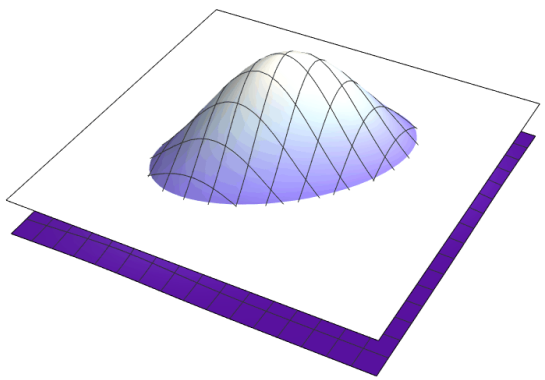


Contour lines

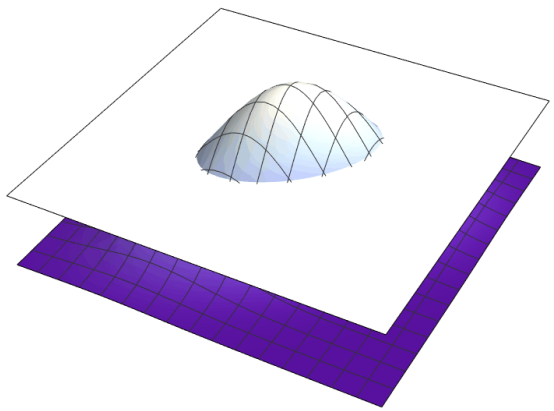


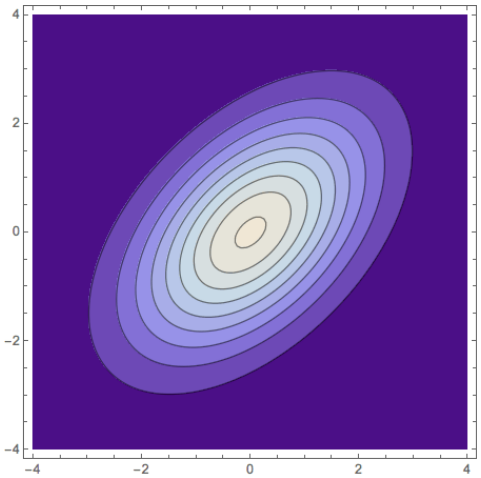
1000 sample points

CONTOUR LINES



Intersect density with a horizontal plane, draw intersection as a curve, and project it down onto the plane.





Each elliptical line is such a contour, for planes at different heights.

Contours and standard deviation

- Each ellipse consists of all points $\mathbf{x} \in \mathbb{R}^2$ that satisfy the equation

$$\left\langle \mathbf{x}, \Sigma^{-1}\mathbf{x} \right\rangle = c \qquad \text{for some fixed } c > 0 \ .$$

  Changing c changes the size of the ellipse.

- The ellipses play the same role as intervals around the mean for 1D Gaussians: The ellipse with $\left\langle \mathbf{x}, \Sigma^{-1}\mathbf{x} \right\rangle = 1$ contains $\sim 68.27\%$ of the probability mass, etc.

- That is: The area within the ellipse given by $\left\langle \mathbf{x}, \Sigma^{-1}\mathbf{x} \right\rangle = k$ corresponds to k standard deviations.

TOOLS: MAXIMUM LIKELIHOOD

Models

A **model** $\mathcal{P}$ is a set of probability distributions. We index each distribution by a parameter value $\theta \in \mathcal{T}$; we can then write the model as

$$\mathcal{P} = \{P_\theta | \theta \in \mathcal{T}\} \ .$$

The set $\mathcal{T}$ is called the **parameter space** of the model.

Parametric model

The model is called **parametric** if the number of parameters (i.e. the dimension of the vector θ) is (1) finite and (2) independent of the number of data points. Intuitively, the complexity of a parametric model does not increase with sample size.

Density representation

For parametric models, we can assume that $\mathcal{T} \subset \mathbb{R}^d$ for some fixed dimension d. We usually represent each P_θ be a density function $p(x|\theta)$.

Setting

- Given: Data $x_1, \ldots, x_n$, parametric model $\mathcal{P} = \{p(x|\theta) \mid \theta \in \mathcal{T}\}$.
- Objective: Find the distribution in $\mathcal{P}$ which best explains the data. That means we have to choose a "best" parameter value $\hat{\theta}$.

Maximum Likelihood approach

Maximum Likelihood assumes that the data is best explained by the distribution in $\mathcal{P}$ under which it has the highest probability (or highest density value).

Hence, the **maximum likelihood estimator** is defined as

$$\hat{\theta}_{\text{ML}} := \arg\max_{\theta \in \mathcal{T}} p(x_1, \ldots, x_n|\theta)$$

the parameter which maximizes the joint density of the data.

The i.i.d. assumption

The standard assumption of ML methods is that the data is **independent and identically distributed (i.i.d.)**, that is, generated by independently sampling repeatedly from the same distrubtion P.

If the density of P is $p(x|\theta)$, that means the joint density decomposes as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

Maximum Likelihood equation

The analytic criterion for a maximum likelihood estimator (under the i.i.d. assumption) is:

$$\nabla_\theta \left(\prod_{i=1}^{n} p(x_i|\theta) \right) = 0$$

We use the "logarithm trick" to avoid a huge product rule computation.

Not examinable.

Recall: Logarithms turn products into sums

$$\log\left(\prod_i f_i\right) = \sum_i \log(f_i)$$

Logarithms and maxima

The logarithm is monotonically increasing on $\mathbb{R}_+$.

Consequence: Application of log does not change the *location* of a maximum or minimum:

$$\max_y \log(g(y)) \neq \max_y g(y) \qquad \text{The } \textit{value} \text{ changes.}$$

$$\arg\max_y \log(g(y)) = \arg\max_y g(y) \qquad \text{The } \textit{location} \text{ does not change.}$$

Not examinable.

Analytic MLE

Likelihood and logarithm trick

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta) = \arg\max_{\theta} \log\Big(\prod_{i=1}^{n} p(x_i|\theta)\Big) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)$$

Analytic maximality criterion

$$0 = \sum_{i=1}^{n} \nabla_\theta \log p(x_i|\theta) = \sum_{i=1}^{n} \frac{\nabla_\theta p(x_i|\theta)}{p(x_i|\theta)}$$

Whether or not we can solve this analytically depends on the choice of the model!

Not examinable.

EXAMPLE: GAUSSIAN MEAN MLE

Model: Multivariate Gaussians

The model $\mathcal{P}$ is the set of all Gaussian densities on $\mathbb{R}^d$ with *fixed* covariance matrix Σ,

$$\mathcal{P} = \{g(\,.\,|\mu, \Sigma) \,|\, \mu \in \mathbb{R}^d\}\ ,$$

where g is the Gaussian density function. The parameter space is $\mathcal{T} = \mathbb{R}^d$.

MLE equation

We have to solve the maximum equation

$$\sum_{i=1}^{n} \nabla_\mu \log g(x_i|\mu, \Sigma) = 0$$

for μ.

Not examinable.

Example: Gaussian Mean MLE

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \nabla_\mu \log \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big) \\
&= \sum_{i=1}^{n} \nabla_\mu \Big(\log\Big(\frac{1}{\sqrt{(2\pi)^d |\Sigma|}}\Big) + \log\Big(\exp\Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big)\Big) \\
&= \sum_{i=1}^{n} \nabla_\mu \Big(-\frac{1}{2}\Big\langle (x_i-\mu), \Sigma^{-1}(x_i-\mu)\Big\rangle\Big) = -\sum_{i=1}^{n} \Sigma^{-1}(x_i-\mu)
\end{aligned}
$$

Multiplication by $(-\Sigma)$ gives

$$
0 = \sum_{i=1}^{n} (x_i - \mu) \qquad \Rightarrow \qquad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i
$$

Conclusion

The maximum likelihood estimator of the Gaussian expectation parameter for fixed covariance is

$$
\hat{\mu}_{\mathrm{ML}} := \frac{1}{n}\sum_{i=1}^{n} x_i
$$

Not examinable.

Model: Multivariate Gaussians

The model $\mathcal{P}$ is now

$$\mathcal{P} = \{g(\,.\,|\mu, \Sigma) \,|\, \mu \in \mathbb{R}^d, \Sigma \in \Delta_d\} \ ,$$

where Δ_d is the set of all possible $d \times d$ covariance matrices. The parameter space is $\mathcal{T} = \mathbb{R}^d \times \Delta_d$.

ML approach

Since we have just seen that the ML estimator of μ does not depend on Σ, we can compute $\hat{\mu}_{\text{ML}}$ first. We then estimate Σ using the criterion

$$\sum_{i=1}^{n} \nabla_\Sigma \log g(x_i|\hat{\mu}_{\text{ML}}, \Sigma) = 0$$

Solution

The ML estimator of Σ is

$$\hat{\Sigma}_{\text{ML}} := \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{\text{ML}})(x_i - \hat{\mu}_{\text{ML}})^t \ .$$

Not examinable.