Multivariate Gaussian

Technische Universität München

reader: Murphy ch. 4

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

$$\boldsymbol{x} \in \mathbb{R}^d, \qquad \boldsymbol{\mu} \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$$

$\boldsymbol{\Sigma}$ is assumed to be symmetric (which means, that $\boldsymbol{\Sigma}^{-1}$ is also symmetric) and positive definite.

$$\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$$
$$\mathrm{Cov}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])^T] = \boldsymbol{\Sigma}$$

Diagonal elements of $\boldsymbol{\Sigma}$?

Expectation is linear. E.g., $\mathbb{E}[\boldsymbol{a}^T\boldsymbol{X}] = \boldsymbol{a}^T\,\mathbb{E}[\boldsymbol{X}]$.

In $2d$, a *bivariate Gaussian* is depicted as an ellipse. Why?

$\boldsymbol{\Sigma}$ is *real* and *symmetric* and therefore

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

$\boldsymbol{U}$ is orthonormal ($\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$).
$\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues.
And thus:

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}^{-T}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T = \sum_{i=1}^{d} \frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T$$

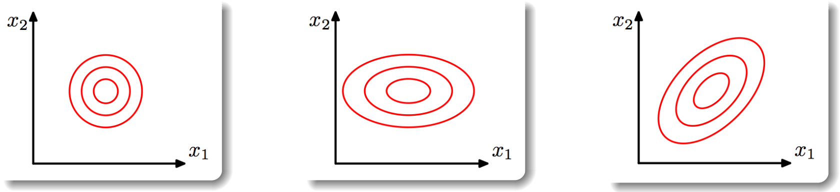
Full rank is assumed.

The *Mahalanobis distance* can be rewritten:

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^T \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \right) (\boldsymbol{x}-\boldsymbol{\mu}) =$$

$$\sum_{i=1}^{D} \frac{1}{\lambda_i} (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{u}_i \underbrace{\boldsymbol{u}_i^T (\boldsymbol{x}-\boldsymbol{\mu})}_{y_i \,\in\, \mathbb{R}} = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

Mahalanobis distance: The Euclidean distance of $\boldsymbol{x}$ from $\boldsymbol{\mu}$ in a rotated and scaled coordinate system.



Linear transformation of a Gaussian:
If

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\xi}$$

then $\boldsymbol{Y}$ is a Gaussian with

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\xi}$$

$$\mathrm{Cov}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

or, in compact notation,

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\xi}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$$

MLE for $\boldsymbol{\mu}$

$$\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$$

Likelihood

$$\prod_{i=1}^n \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and thus the *negative* log-likelihood is

$$\underbrace{\frac{nd}{2}\log 2\pi}_{\text{const.}} + \underbrace{\frac{n}{2}\log|\boldsymbol{\Sigma}|}_{\text{depends on } \boldsymbol{\Sigma}} + \underbrace{\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}_{\text{depends on } \boldsymbol{\mu}, \boldsymbol{\Sigma}} := \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

We need the derivative w.r.t. $\boldsymbol{\mu}$, and therefore first consider:

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

The derivative of this term w.r.t. $\boldsymbol{\mu}$ is (helpful: $\frac{\partial \boldsymbol{a}^T \boldsymbol{y}}{\partial \boldsymbol{y}} = \boldsymbol{a}$, $\frac{\partial \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}}{\partial \boldsymbol{y}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{y}$)

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

So

$$\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \frac{1}{2} \sum_{i=1}^{n} 2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{x}_i) = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{\mu} - \boldsymbol{x}_i)$$

Optimum at

$$\boldsymbol{\mu}^*_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

(second derivative at $\boldsymbol{\mu}_{\mathrm{MLE}}$ is $2\boldsymbol{\Sigma}^{-1}$)

The trace operator: $tr(\boldsymbol{A}) := \sum_i \boldsymbol{A}_{ii}$ has a *cyclic* property

$$tr(\boldsymbol{ABC}) = tr(\boldsymbol{BCA}) = tr(\boldsymbol{CAB})$$

that allows the re-casting:

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = tr(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}) = tr(\boldsymbol{A} \boldsymbol{x} \boldsymbol{x}^T)$$

i.e.

$$tr((\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})) = tr(\boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T)$$

One more fact:

$$|\boldsymbol{\Sigma}| = \frac{1}{|\boldsymbol{\Sigma}^{-1}|}$$

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = \text{const.} - \frac{n}{2} \log |\boldsymbol{\Sigma}^{-1}| + \frac{1}{2} \sum_i tr(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T)$$

$\left(\frac{\partial \log |\boldsymbol{A}|}{\partial \boldsymbol{A}} = \boldsymbol{A}^{-T}, \ \frac{\partial tr(\boldsymbol{BA})}{\partial \boldsymbol{A}} = \boldsymbol{B}^T \right)$

$$\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})}{\partial \boldsymbol{\Sigma}^{-1}} = -\frac{n}{2} \boldsymbol{\Sigma}^T + \frac{1}{2} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Extremum at

$$\boldsymbol{\Sigma}^*_{\text{MLE}} = \frac{1}{n} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Central Limit Theorem

Let $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ be i.i.d. random variables with finite mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$, then

$$\boldsymbol{S}_n := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i - \boldsymbol{\mu}\right) \Rightarrow \boldsymbol{S}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

if you average i.i.d. variables, then only mean and covariance are retained (everything else is smoothed away) and a Gaussian remains.

Maximum entropy distributions

How much uncertainty is in a distribution?

Differential entropy for continuous distribution P:

$$\mathcal{H}[P] = \int p(\boldsymbol{x})(-\log p(\boldsymbol{x}))\boldsymbol{dx}$$

Given mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ (and nothing else!), what distribution has highest differential entropy?

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \arg\max_P(\mathcal{H}[P] \mid \mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}, \mathrm{Cov}[\boldsymbol{X}] = \boldsymbol{\Sigma})$$

Upper bound on entropy:

$$\frac{1}{2}\log((2\pi e)^D|\boldsymbol{\Sigma}|)$$

How to sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\boldsymbol{\Sigma}$ is real, symmetric and positive definite. Thus, *Cholesky decomposition* exists:

$$\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$$

where $\boldsymbol{L}$ is lower triangular with strictly positive diagonal entries.

If we can easily sample **Z** from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (we usually can, why?), i.e. $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, then

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z} \Rightarrow \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Why?

How to evaluate $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Evaluate *density* at some point $\boldsymbol{x}$, e.g. to compute log likelihood.

We need to compute $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^{-T} \boldsymbol{L}^{-1}$$

and compute in a *numerically stable* way

$$\boldsymbol{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

Also

$$|\boldsymbol{\Sigma}| = \prod_{j=1}^{D} L_{jj}^2$$

(prefer to work in the log domain!)

Normalisation factor

There is the elementary result for the 1d normal distribution:

$$\int e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \sqrt{2\pi\sigma^2}$$

Remember the eigendecomposition of the covariance matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

Denote by

$$f(\boldsymbol{x}) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

the unnormalised probability density function of $\boldsymbol{x}$.
We need to compute

$$\int f(\boldsymbol{x})\boldsymbol{dx}$$

Normalization factor

Defining $\boldsymbol{y} = \boldsymbol{U}^T(\boldsymbol{x} - \boldsymbol{\mu})$, the integral changes as follows (*change of variable*):

$$\int f(\boldsymbol{x})\boldsymbol{dx} = \int f(\boldsymbol{x(y)}) \left| \frac{\boldsymbol{dx}}{\boldsymbol{dy}} \right| \boldsymbol{dy} = \int f(\boldsymbol{x(y)}) |\boldsymbol{U}| \boldsymbol{dy} = \int \prod_i e^{-\frac{y_i^2}{2\lambda_i}} \boldsymbol{dy}$$

Using the elementary result:

$$\int \prod_i e^{-\frac{y_i^2}{2\lambda_i}} \boldsymbol{dy} = \prod_i \int e^{-\frac{y_i^2}{2\lambda_i}} dy_i = \prod_i \sqrt{2\pi\lambda_i} = \sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}$$

Interesting Observation:

$$\text{diagonal } \boldsymbol{\Sigma} \Rightarrow p(\boldsymbol{x}) = \prod_i p(x_i)$$

In words: For Gaussians, uncorrelated components induce independent components (what is the *general* rule?).

Products of Gaussians

What is the product of two Gaussian pdfs?

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

Difficult to answer in *moment parameterisation* form:

$$\propto e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Natural (or: canonical) parameterisation:

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}+\boldsymbol{r}^T\boldsymbol{x}}, \qquad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1}, \boldsymbol{r} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

($\propto$ ignores all constants (i.e., terms without $\boldsymbol{x}$). Do the transformation to natural parameters on your own!)

A Gaussian pdf is written in *information form* (or: exponential family form), if natural parameterisation is used.

Products of Gaussians

Using natural parameters, we can write:

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \propto e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{r}_1^T \boldsymbol{x}} \cdot e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}_2 \boldsymbol{x} + \boldsymbol{r}_2^T \boldsymbol{x}}$$

The result is *again* a Gaussian, because we can write it in information form:

$$\propto e^{-\frac{1}{2}\boldsymbol{x}^T (\boldsymbol{A}_1 + \boldsymbol{A}_2) \boldsymbol{x} + (\boldsymbol{r}_1 + \boldsymbol{r}_2)^T \boldsymbol{x}}$$

Converting it back into moment parameterisation gives a new $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}, \qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu_1} + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu_2})$$

Now we can compute the missing normalisation constant for the resulting Gaussian.

(do back-transformations on your own)

Marginalisation

$$I \subset \{1, 2, \ldots, d\}, \quad \boldsymbol{X}_I := (X_i)_{i \in I}$$

What is $p(\boldsymbol{x}_I)$?

Linear Transformation with a *selection matrix*:

$$\boldsymbol{X}_I = \boldsymbol{I}_I \boldsymbol{X}$$

That is, $\boldsymbol{X}_I$ is Gaussian:

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

Conditioning

$$I \subset \{1, 2, \dots, d\}, \quad R = \{1, 2, \dots, d\} \setminus I$$

What is $p(\boldsymbol{x}_I \mid \boldsymbol{x}_R)$?

A Gaussian, but what does it look like?
Rather straightforward when using natural parameterisation by remembering

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_I \mid \boldsymbol{x}_R) p(\boldsymbol{x}_R)$$

Basically, read result after some algebraic reformulations.
Not immediately obvious, if we want the results in moment parametrisation (needs *Schur* complement for inverting partitioned matrices). Thus, just the results:

$$\boldsymbol{\mu}_{I|R} = \boldsymbol{\mu}_I + \boldsymbol{\Sigma}_{IR} \boldsymbol{\Sigma}_{RR}^{-1} (\boldsymbol{x}_R - \boldsymbol{\mu}_R)$$

$$\boldsymbol{\Sigma}_{I|R} = \boldsymbol{\Sigma}_{II} - \boldsymbol{\Sigma}_{IR} \boldsymbol{\Sigma}_{RR}^{-1} \boldsymbol{\Sigma}_{RI}$$

Linear Gaussian systems

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

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_{Y|X})$$

($\boldsymbol{x}, \boldsymbol{y}$ can have different dimensionalities)

What is $p(\boldsymbol{x} \mid \boldsymbol{y})$ and $p(\boldsymbol{y})$?

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{X|Y}, \boldsymbol{\Sigma}_{X|Y})$$

$$\boldsymbol{\Sigma}_{X|Y} = (\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{A}^T \boldsymbol{\Sigma}_{Y|X}^{-1} \boldsymbol{A})^{-1}$$

$$\boldsymbol{\mu}_{X|Y} = \boldsymbol{\Sigma}_{X|Y} \left(\boldsymbol{A}^T \boldsymbol{\Sigma}_{Y|X}^{-1} (\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X \right)$$

Condition x on a noisy observation of itself.

Let $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})^T$:
By observing that $p(\boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$ and properly rearranging we find (see Bishop section 2.3.3):

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$$

$$\boldsymbol{\mu}_Z = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{A}\boldsymbol{\mu}_X + \boldsymbol{b} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_Z = \begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_X \boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{Y|X} + \boldsymbol{A}\boldsymbol{\Sigma}_X \boldsymbol{A}^T \end{pmatrix}$$

And finally, $\boldsymbol{Y}$:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu}_X + \boldsymbol{b}, \boldsymbol{\Sigma}_{Y|X} + \boldsymbol{A}\boldsymbol{\Sigma}_X \boldsymbol{A}^T)$$

Inferring an unknown vector from noisy measurements

Assume, that $\boldsymbol{X}$ represents the true, but unknown location of some object (e.g. could be 2d/3d position). Model this by

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma}_0)$$

We make noisy observations $\boldsymbol{Y}_i$ of $\boldsymbol{X}$:

$$\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{X}, \boldsymbol{\Sigma}_Y)$$

This means, we know in what way our sensor errs.
Compared to the general form above, $\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{b} = \boldsymbol{0}$.

$$p(\boldsymbol{x} \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

$$\boldsymbol{\Sigma}_n = (\boldsymbol{\Sigma_0}^{-1} + n\boldsymbol{\Sigma}_y^{-1})^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_n\Big(\boldsymbol{\Sigma}_y^{-1}\Big(\sum_i \boldsymbol{y}_i\Big) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\Big)$$

You can use the same idea to do sensor fusion (different kinds of sensors with different kinds of measure noise).

Bayes for Gaussian

With the previous formulae, we finally can do a Bayesian approach for Gaussians. To simplify the derivation, we only consider the case $p(\boldsymbol{\mu} \mid \mathcal{D}, \boldsymbol{\Sigma})$.

That is, we want to determine the posterior distribution for $\boldsymbol{\mu}$ from observations $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, where we assume that the covariance $\boldsymbol{\Sigma}$ of these observations is known.

Gaussian prior for $\boldsymbol{\mu}$:

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$$

Then

$$p(\boldsymbol{\mu} \mid \mathcal{D}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{V}_n)$$

$$\boldsymbol{V}_n = (\boldsymbol{V_0}^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{V}_n\Big(\boldsymbol{\Sigma}^{-1}\Big(\sum_i \boldsymbol{x}_i\Big) + \boldsymbol{V}_0^{-1}\boldsymbol{\mu}_0\Big)$$

Bayes for Gaussian

If we don't know anything about the prior (uninformative prior, i.e., $\boldsymbol{V}_0^{-1} = 0\boldsymbol{I}$), then

$$p(\boldsymbol{\mu} \mid \mathcal{D}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\frac{1}{n}\sum_i \boldsymbol{x}_i, \frac{1}{n}\boldsymbol{\Sigma}\right)$$

Remember the MLE of $\boldsymbol{\mu}$?