Maximising likelihood for multivariate Gaussian distributions

When you find the mean of some data, what you are really doing is finding a parameter which *maximises the likelihood*.

For instance, assume some points $\{\vec{x}_1, ..., \vec{x}_N\} \subset \mathbb{R}^d$ are normally distributed. The conditional probability of the data is

$$P\left(\vec{x}_i \mid \vec{\mu}, \Sigma\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{\tau^d \det \Sigma}} \exp\left(-\frac{1}{2}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)\right)$$

which is also the *likelihood* of the parameters $\vec{\mu}$ and Σ. It is often easiler to manipulate the logarithm of the likelihood:

$$\log P = -\frac{1}{2}\log\left(\tau^d \det \Sigma\right) - \frac{1}{2}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)$$

Fitting $\vec{\mu}$ to data

To find the mean $\vec{\mu}$ which maximises the likelihood, note that $\log P$ is quadratic in $\vec{\mu}$, so the maximum occurs at the unique point where its derivative vanishes.

Consider the differential $\delta \log P$ induced by $\vec{\mu} \to \vec{\mu} + \delta\vec{\mu}$:

$$\begin{aligned}
\delta \log P &= \frac{1}{2}\sum_{i=1}^{N}\left[\left(\delta\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right) + \left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\delta\vec{\mu}\right] \\
&= \sum_{i=1}^{N}\left[\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\right]\delta\vec{\mu} \\
&= \left[\sum_{i=1}^{N}\vec{x}_i - N\vec{\mu}\right]^T \Sigma^{-1}\delta\vec{\mu}
\end{aligned}$$

If the likelihood is at a local maximum, then $\delta \log P$ must vanish for any $\delta\vec{\mu}$. This holds when:

$$\boxed{\vec{\mu} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}_i}$$

Fitting Σ to data

To find the covariance matrix Σ which maximises the likelihood, consider the differential likelihood induced by $\Sigma \to \Sigma + \delta\Sigma$.

$$\delta \log P = -\frac{N}{2}\frac{\delta \det \Sigma}{\det \Sigma} - \frac{1}{2}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)^T \delta\left(\Sigma^{-1}\right)\left(\vec{x}_i - \vec{\mu}\right)$$

Use the identities

$$\begin{aligned}
\delta \det A &= \operatorname{tr}\left[A^{-1}\delta A\right]\det A \\
\delta\left(\Sigma^{-1}\right) &= \Sigma^{-2}\delta\Sigma = \delta\Sigma\,\Sigma^{-2}
\end{aligned}$$

and take the trace to obtain

$$\operatorname{tr}[\delta \log P] = -\frac{N}{2}\operatorname{tr}\left[\Sigma^{-1}\delta\Sigma\right] - \frac{1}{2}\sum_{i=1}^{N}\operatorname{tr}\left[\left(\vec{x} - \vec{\mu}\right)\left(\vec{x} - \vec{\mu}\right)^T \Sigma^{-2}\delta\Sigma\right]$$

where we use the cyclic property of the trace in the last term.

If the likelihood is at a local maximum, then it vanishes for any $\delta\Sigma$. Since $\delta\Sigma$ is arbitrary, this scalar equality between trace implies equality between the matrices themselves:

$$-2\delta \log P = -N\Sigma^{-1}\delta\Sigma + \sum_{i=1}^{N}\left(\vec{x} - \vec{\mu}\right)\left(\vec{x} - \vec{\mu}\right)^T \Sigma^{-2}\delta\Sigma$$

This vanishes when the covariance matrix is given by:

$$\boxed{\Sigma = \frac{1}{N}\sum_{i=1}^{N}\left(\vec{x} - \vec{\mu}\right)\left(\vec{x} - \vec{\mu}\right)^T}$$