MLE of multivariate Gaussian

CS5014

Lei Fang

In this note, we derive the maximum likelihood estimator for multivariate Gaussian. Given $\{x^1, x^2, \ldots, x^n\}$, assume $x^i \sim N(\mu, \Sigma)$; find the ML estimate of

$$\mu,\ \Sigma$$

The log likelihood is:

$$\begin{aligned}
\mathcal{L}(\mu, \Sigma) = \log p(\{x^i\}_1^n | \mu, \Sigma) &= \sum_{i=1}^n \log N(x^i; \mu, \Sigma) \\
&= \sum_{i=1}^n -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu) - \frac{d}{2}\log 2\pi \\
&= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n (x^i - \mu)^\top \Sigma^{-1}(x^i - \mu) + \text{const.}
\end{aligned}$$

Now ready to compute the ML estimator

MLE for $\hat{\mu}$

$$\mathcal{L}(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu) + \text{const.}$$

The MLE are defined as usual:

$$\hat{\mu}, \hat{\Sigma} = \arg\max_{\mu, \Sigma} \mathcal{L}(\mu, \Sigma)$$

Remember I have mentioned that quadratic form is multivariate generalisation of quadratic function:

$$x^\top A x$$

is simular to $x \cdot a \cdot x$; their gradients are similar as well

$$\frac{\partial xax}{\partial x} = \frac{\partial ax^2}{\partial x} = (a + a)x = 2ax$$

The gradient of quadratic form is

$$\frac{\partial x^\top A x}{\partial x} = (A + A^\top)x = 2Ax$$

if we assume A is symmetric, then $A^\top + A = 2A$.

Take derivative w.r.t μ (notice it is a quadratic form w.r.t μ) and set it to zero:

$$\nabla_\mu \mathcal{L} = -\frac{1}{2}\sum_{i=1}^{n} 2 \cdot (-1) \cdot \Sigma^{-1}(x^i - \mu) = \sum_{i=1}^{n}\Sigma^{-1}(x^i - \mu) = 0$$

which leads to

$$\begin{aligned}
&\Rightarrow \Sigma^{-1}\sum_{i=1}^{n}(x^i - \mu) = 0 \\
&\Rightarrow \sum_{i}^{n}(x^i - \mu) = 0 \\
&\Rightarrow n \cdot \mu = \sum_{i=1}^{n} x^i \\
&\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^i
\end{aligned}$$

where

- the first step left multiplies Σ on both sides, $\Sigma\Sigma^{-1} = I$

MLE for $\hat{\Sigma}$

The same deal, we need to take derivative and then set the derivative to zero.

We will use the trace trick here first to rewrite the log likelihood function;

Note that

$$\text{Trace}(X) = \sum_i X_{ii},$$

the sum of the diagonal entries of a matrix; and also $\text{Trace}(c) = c$, i.e. trace of a scalar is itself; and trace is a linear operator,

$$\text{Trace}\left(\sum_{i=1}^n w_i A_i\right) = \sum_{i=1}^n w_i \text{Trace}(A_i),$$

and trace has a cyclic property:

$$\text{Trace}(ABC) = \text{Trace}(CAB) = \text{Trace}(BCA)$$

We are ready to rewrite the log likelihood now:

$$\begin{aligned}
\mathcal{L}(\mu, \Sigma) &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n (x^i - \mu)^\top \Sigma^{-1}(x^i - \mu) + \text{const.} \\
&= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n \text{Trace}\big((x^i - \mu)^\top \Sigma^{-1}(x^i - \mu)\big) + \text{const.} \\
&= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n \text{Trace}\big(\Sigma^{-1}(x^i - \mu)(x^i - \mu)^\top\big) + \text{const.} \\
&= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\text{Trace}\left(\Sigma^{-1}\sum_{i=1}^n (x^i - \mu)(x^i - \mu)^\top\right) + \text{const.}
\end{aligned}$$

where

- line two has used the property that trace of a scalar is itself; a quadratic form is a scalar; as well as linear operator
- line three: cyclic property
- line four: linear operator again

Check matrix cookbook for some useful matrix gradient identities:

$$\frac{\partial \ln|X|}{\partial X} = (X^{-1})^\top, \quad \frac{\partial \text{Trace}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})^\top$$

Then the gradient becomes:

$$\nabla_\Sigma \mathcal{L} = -\frac{n}{2}\left(\Sigma^{-1}\right)^\top - \frac{1}{2}\left[-\left(\Sigma^{-1}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top\Sigma^{-1}\right)^\top\right]$$

Set the derivative to zero, we have

$$\begin{aligned}
&-\frac{n}{2}\left(\Sigma^{-1}\right)^\top - \frac{1}{2}\left[-\left(\Sigma^{-1}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top\Sigma^{-1}\right)^\top\right] = 0 \\
&\Rightarrow n\cdot\Sigma^{-1} = \left(\Sigma^{-1}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top\Sigma^{-1}\right) \\
&\Rightarrow \Sigma\cdot n = \sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top \\
&\Rightarrow \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^\top
\end{aligned}$$

where

- the first step multiplies -2 on both side and move the second term to the right hand side of the equation; and then take transpose on both side: i.e. $(A^\top)^\top = A$
- the second step multplies both left and right hand side with Σ, note $\Sigma^{-1}\Sigma = \Sigma\Sigma^{-1} = I$
- the third step does not need explanation...

So MLE for multivariate Gaussians are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x^i; \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x^i-\hat{\mu})(x^i-\hat{\mu})^\top$$

- intuitive results: empirical sample mean and covariance are the estimators!

Weighted MLE should be very similar. Typing latex is very painful. I will stop here and leave it as an exercise...