MLE for Multivariate Gaussian Distribution

Assume we have N examples from training data (iid) and the likelihood function is:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i-1}^{N} \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{Nn}{2} \log 2\pi - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

Taking derivatives w.r.t. μ, we have

$$\frac{\partial}{\partial \boldsymbol{\mu}} \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}} \left(\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) = \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}} \left(\sum_{i=1}^{N} 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) \right)$$

$$= \sum_{i=1}^{N} (\Sigma^{-1} \mathbf{x}_i - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu})$$

Setting it to zero we obtain: $\sum_{i=1}^{N} \Sigma^{-1} \mathbf{x}_i = N \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$

Thus the ML estimate of $\boldsymbol{\mu}$ is:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

To estimate Σ, we take derivative w.r.t. Σ:

$$\frac{\partial}{\partial \Sigma} \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2} \frac{\partial}{\partial \Sigma} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \frac{\partial}{\partial \Sigma} \left(\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)$$

For the first term, we have:

$$\frac{\partial}{\partial \Sigma} \log|\boldsymbol{\Sigma}| = \frac{1}{|\boldsymbol{\Sigma}|} \frac{\partial}{\partial \Sigma} (|\boldsymbol{\Sigma}|) = \frac{1}{|\boldsymbol{\Sigma}|} |\boldsymbol{\Sigma}| \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}, \quad \left(\text{note that } \frac{\partial}{\partial \Sigma} |\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}| \boldsymbol{\Sigma}^{-1}, \text{ see } \textit{Matrix Cookbook}\right)$$

For the second term, we use the following result from the *Matrix Cookbook*:

$$\frac{\partial}{\partial \Sigma} \mathbf{a}^T \Sigma^{-1} \mathbf{b} = -\boldsymbol{\Sigma}^{-1} \mathbf{a} \mathbf{b}^{\mathbf{T}} \boldsymbol{\Sigma}^{-1}$$

Therefore

$$\frac{\partial}{\partial \Sigma} \left(\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) = -\sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{\mathbf{i}} - \boldsymbol{\mu})(\mathbf{x}_{\mathbf{i}} - \boldsymbol{\mu})^{\mathbf{T}} \boldsymbol{\Sigma}^{-1}$$

It follows that:

$$\frac{\partial}{\partial \Sigma} \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2} \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}$$

Set the derivative to zero and replace μ with their ML estimate, and after a few steps of derivations, we obtain the ML estimate of $\boldsymbol{\Sigma}$:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$