4.1.3 MLE for an MVN

We now describe one way to estimate the parameters of an MVN, using MLE. In later sections, we will discuss Bayesian inference for the parameters, which can mitigate overfitting, and can provide a measure of confidence in our estimates.

Theorem 4.1.1 (MLE for a Gaussian). *If we have N iid samples $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE for the parameters is given by*

$$\hat{\boldsymbol{\mu}}_{mle} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \triangleq \overline{\mathbf{x}} \tag{4.6}$$

$$\hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T) - \overline{\mathbf{x}}\,\overline{\mathbf{x}}^T \tag{4.7}$$

That is, the MLE is just the empirical mean and empirical covariance. In the univariate case, we get the following familiar results:

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \overline{x} \tag{4.8}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_i x_i^2\right) - (\overline{x})^2 \tag{4.9}$$

4.1.3.1 Proof *

To prove this result, we will need several results from matrix algebra, which we summarize below. In the equations, $\mathbf{a}$ and $\mathbf{b}$ are vectors, and $\mathbf{A}$ and $\mathbf{B}$ are matrices. Also, the notation $\text{tr}(\mathbf{A})$ refers to the **trace** of a matrix, which is the sum of its diagonals: $\text{tr}(\mathbf{A}) = \sum_i A_{ii}$.

$$\boxed{\begin{aligned}
\frac{\partial(\mathbf{b}^T\mathbf{a})}{\partial \mathbf{a}} &= \mathbf{b} \\
\frac{\partial(\mathbf{a}^T\mathbf{A}\mathbf{a})}{\partial \mathbf{a}} &= (\mathbf{A} + \mathbf{A}^T)\mathbf{a} \\
\frac{\partial}{\partial \mathbf{A}}\text{tr}(\mathbf{B}\mathbf{A}) &= \mathbf{B}^T \\
\frac{\partial}{\partial \mathbf{A}}\log|\mathbf{A}| &= \mathbf{A}^{-T} \triangleq (\mathbf{A}^{-1})^T \\
\text{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) &= \text{tr}(\mathbf{C}\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{C}\mathbf{A})
\end{aligned}} \tag{4.10}$$

The last equation is called the **cyclic permutation property** of the trace operator. Using this, we can derive the widely used **trace trick**, which reorders the scalar inner product $\mathbf{x}^T\mathbf{A}\mathbf{x}$ as follows

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \text{tr}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{x}\mathbf{x}^T\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^T) \tag{4.11}$$

Proof. We can now begin with the proof. The log-likelihood (dropping additive constants) is given by

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{4.12}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix.

Using the substitution $\mathbf{y}_i = \mathbf{x}_i - \boldsymbol{\mu}$ and the chain rule of calculus, we have

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\mu}}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) &= \frac{\partial}{\partial\mathbf{y}_i}\mathbf{y}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{y}_i\frac{\partial\mathbf{y}_i}{\partial\boldsymbol{\mu}} && (4.13)\\ &= -1(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-T})\mathbf{y}_i && (4.14) \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\mu}}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\sum_{i=1}^{N} -2\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}) = 0 && (4.15)\\ \hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \overline{\mathbf{x}} && (4.16) \end{aligned}$$

So the MLE of $\boldsymbol{\mu}$ is just the empirical mean.

Now we can use the trace-trick to rewrite the log-likelihood for $\boldsymbol{\Lambda}$ as follows:

$$\begin{aligned} \ell(\boldsymbol{\Lambda}) &= \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i}\mathrm{tr}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Lambda}] && (4.17)\\ &= \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\mathrm{tr}\left[\mathbf{S}_{\mu}\boldsymbol{\Lambda}\right] && (4.18)\\ & && (4.19) \end{aligned}$$

where

$$\mathbf{S}_{\mu} \triangleq \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{4.20}$$

is the scatter matrix centered on $\boldsymbol{\mu}$. Taking derivatives of this expression with respect to $\boldsymbol{\Lambda}$ yields

$$\begin{aligned} \frac{\partial\ell(\boldsymbol{\Lambda})}{\partial\boldsymbol{\Lambda}} &= \frac{N}{2}\boldsymbol{\Lambda}^{-T} - \frac{1}{2}\mathbf{S}_{\mu}^T = 0 && (4.21)\\ \boldsymbol{\Lambda}^{-T} &= \boldsymbol{\Lambda}^{-1} = \boldsymbol{\Sigma} = \frac{1}{N}\mathbf{S}_{\mu} && (4.22) \end{aligned}$$

so

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \tag{4.23}$$

which is just the empirical covariance matrix centered on $\boldsymbol{\mu}$. If we plug-in the MLE $\boldsymbol{\mu} = \overline{\mathbf{x}}$ (since both parameters must be simultaneously optimized), we get the standard equation for the MLE of a covariance matrix. □

4.1.4 Maximum entropy derivation of the Gaussian *

In this section, we show that the multivariate Gaussian is the distribution with maximum entropy subject to having a specified mean and covariance (see also Section 9.2.6). This is one reason the Gaussian is so widely used: the first two moments are usually all that we can reliably estimate from data, so we want a distribution that captures these properties, but otherwise makes as few additional assumptions as possible.

To simplify notation, we will assume the mean is zero. The pdf has the form

$$p(\mathbf{x}) = \frac{1}{Z}\exp(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}) \tag{4.24}$$

If we define $f_{ij}(\mathbf{x}) = x_i x_j$ and $\lambda_{ij} = \frac{1}{2}(\Sigma^{-1})_{ij}$, for $i, j \in \{1, \ldots, D\}$, we see that this is in the same form as Equation 9.74. The (differential) entropy of this distribution (using log base e) is given by

$$h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2}\ln\left[(2\pi e)^D|\boldsymbol{\Sigma}|\right] \tag{4.25}$$

We now show the MVN has maximum entropy amongst all distributions with a specified covariance $\boldsymbol{\Sigma}$.

Theorem 4.1.2. *Let $q(\mathbf{x})$ be any density satisfying $\int q(\mathbf{x}) x_i x_j d\mathbf{x} = \Sigma_{ij}$. Let $p = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then $h(q) \leq h(p)$.*

Proof. (From (Cover and Thomas 1991, p234).) We have

$$\begin{aligned} 0 &\leq \mathbb{KL}(q||p) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} & (4.26)\\ &= -h(q) - \int q(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} & (4.27)\\ &=^* -h(q) - \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} & (4.28)\\ &= -h(q) + h(p) & (4.29) \end{aligned}$$

where the key step in Equation 4.28 (marked with a *) follows since q and p yield the same moments for the quadratic form encoded by $\log p(\mathbf{x})$. □

4.2 Gaussian discriminant analysis

One important application of MVNs is to define the class conditional densities in a generative classifier, i.e.,

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{4.30}$$