Max. Likelihood Estimate for Multivariate Gaussian

Ashim Gupta
ashimgupta95@gmail.com

1 Introduction

Maximum Likelihood Estimation (MLE) is a popular method in machine learning for estimating parameters of a statistical model. A widely known result for a multivariate Gaussian is that the maximum likelihood estimate for a Gaussian yields an *unbiased* estimate for the mean but a *biased* estimate for the covariance. In these notes, we will look at its proof multivariate case.

Note, that this essentially forms the proof for Exercise 2.35 in PRML book.

2 Biased/Unbiased Maximum Likelihood Estimates

As in Bishop and Nasrabadi [2006], we will assume that we are given N observations, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, with each $\mathbf{x}_n$ being d dimensional, and drawn independently from a multivariate Gaussian distribution. From the book, the maximum likelihood estimates for mean ($\boldsymbol{\mu}$) and covariance ($\boldsymbol{\Sigma}$) are:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{1}$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T \tag{2}$$

These can derived by setting the first derivatives of log-likelihood equal to zero for each of the two parameters. In the book, the expectation for the maximum likelihood solutions of the two parameters is given as:

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu} \tag{3}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N}\boldsymbol{\Sigma} \tag{4}$$

Now, since the expectation of the estimate, $\boldsymbol{\mu}_{ML}$, is equal to the true mean ($\boldsymbol{\mu}$), this estimate is *unbiased*. And because expectation of $\boldsymbol{\Sigma}_{ML}$ is not equal to its true value, that estimate is *biased*, or that the covariance parameter is underestimated. We will see why this is the case.

ML estimate for mean is unbiased. We take the expectation in eq. 1 on both sides:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{ML}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathbf{x}_n] = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\mu} = \frac{1}{N}N\boldsymbol{\mu} = \boldsymbol{\mu}\end{aligned}$$

which gives the equation 3. We made use of linearity of expectation in this short proof.

ML estimate for covariance is biased.

$$\mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T\right] \tag{5}$$

In order to prove this, we will first prove a few other simpler results.

1.

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \tag{6}$$

Proof: From the definition of $\boldsymbol{\Sigma}$,

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \mathbb{E}(\mathbf{x}\mathbf{x}^T - \mathbf{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T)\\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

Simplifying further gives the required result.

2.

$$\mathbb{E}[\mathbf{x}_i\mathbf{x}_j^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T, i \neq j \tag{7}$$

Proof: This follows the same technique as above. Since, we have an *iid* assumption, the co-variance of two different data points $\mathbf{x}_i, \mathbf{x}_j$ should be zero.

$$\begin{aligned}0 &= \mathbb{E}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T] = \mathbb{E}(\mathbf{x}_i\mathbf{x}_j^T - \mathbf{x}_i\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{x}_j^T + \boldsymbol{\mu}\boldsymbol{\mu}^T)\\ &= \mathbb{E}[\mathbf{x}_i\mathbf{x}_j^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

which after simplifying gives the required result.

Now, let us return to eq 5. We will simply expand the right hand side of the equation.

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Sigma}_{ML}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T\right]\\ &= \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}(\mathbf{x}_n\mathbf{x}_n^T - \mathbf{x}_n\boldsymbol{\mu}_{ML}^T - \boldsymbol{\mu}_{ML}\mathbf{x}_n^T + \boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T)\right]\end{aligned}$$

We will look at these four terms separately:

$$\mathrm{I} = \mathbb{E}\left[\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\right] = \sum_{n=1}^{N}\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^T] = N * (\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}) \tag{Using 6}$$

$$\mathrm{II} = \mathbb{E}\left[\sum_{n=1}^{N}\mathbf{x}_n\boldsymbol{\mu}_{ML}^T\right] = \mathbb{E}[N * \boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T] = N * \mathbb{E}[\boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T]$$

$$\text{III} = \mathbb{E}\left[\sum_{n=1}^{N} \boldsymbol{\mu}_{ML}\mathbf{x}_n^T\right] = \mathbb{E}[N * \boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T] = N * \mathbb{E}[\boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T]$$

$$\text{IV} = \mathbb{E}\left[\sum_{n=1}^{N} \boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T\right] = N * \mathbb{E}[\boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T]$$

The only quantity left to compute is $\mathbb{E}[\boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T]$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{ML}\boldsymbol{\mu}_{ML}^T] &= \mathbb{E}\left[(\sum_{i=1}^{N} \frac{1}{N}\mathbf{x}_i)(\sum_{j=1}^{N} \frac{1}{N}\mathbf{x}_j)^T\right] = \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_j^T\right] \\
&= \frac{1}{N^2}(N^2 * \boldsymbol{\mu}\boldsymbol{\mu}^T + N * \boldsymbol{\Sigma}) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma} \qquad \text{(Using results from 6, 7)}
\end{aligned}$$

Finally, we substitute these values in the equation for $\mathbb{E}[\boldsymbol{\Sigma}_{ML}]$

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma}_{ML}] &= \frac{1}{N}\left(N * (\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}) - N * (\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma})\right) \\
\mathbb{E}[\boldsymbol{\Sigma}_{ML}] &= \frac{1}{N}(N\boldsymbol{\Sigma} - \boldsymbol{\Sigma}) = \frac{N-1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

References

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.