4.12

Intro

In this exercise we will study a very popular criteria for model selection. It is called Bayesian Information Criterion (BIC) and its main idea its to penalize complex models so that they do not overfit the data. The BIC is expressed as: $BIC = logp(D|\theta_{ML}) - \frac{d}{2}log(N)$, a combination of the maximum likelihood and a penalization term. The latter term penalize free parameters (d), so that we don't artificially increase the likelihood score by adding new terms to our model. In another words, the BIC draw a line where we should stop adding new parameters to increase the likelihood.

In this particular exercise, we will use BIC to compare two opposite Gaussian models. The first is a MVN with a full covariance matrix, while the second is a MVN with a diagonal matrix. The BIC will help us to determine if it is worth to use the former complex model or the latter more simple approach.

Solution

a)

The BIC for the full covariance MVN is:

$$BIC = -\frac{N}{2}tr(\hat{\Sigma}_{full}^{-1}\hat{S}) - \frac{N}{2}log(|\hat{\Sigma}_{full}|) - \frac{d}{2}log(N) \tag{1}$$

Since we are working with the MLE, $\hat{\Sigma}_{full} = \hat{S}$. Therefore:

$$\begin{aligned} BIC = -\frac{N}{2}tr(I_D) - \frac{N}{2}log(|\hat{\Sigma}_{full}|) - \frac{d}{2}log(N) = \\ -\frac{N}{2}D - \frac{N}{2}log(|\hat{\Sigma}_{full}|) - \frac{d}{2}log(N) \end{aligned} \tag{2}$$

Now, we need to determine the number of free parameters of the model. We have D parameters for the mean and $\frac{D(D+1)}{2}$ parameters for the covariance matrix, since it is full and symmetric. Substituting this in the BIC, we arrive at:

$$\begin{aligned} BIC = -\frac{N}{2}D - \frac{N}{2}log(|\hat{\Sigma}_{full}|) - \frac{\frac{D(D+1)}{2} + D}{2}log(N) = \\ -\frac{N}{2}log(|\hat{\Sigma}_{full}|) - \frac{D^2 + 3D}{4}log(N) - \frac{N}{2}D \end{aligned} \tag{3}$$

b)

The procedure for the MVN with the diagonal matrix is exactly the same. The difference is in the number of free parameters on the covariance matrix, which is in this case equal to D. So now we have:

$$\begin{aligned} BIC = -\frac{N}{2}D - \frac{N}{2}log(|\hat{\Sigma}_{diag}|) - \frac{D + D}{2}log(N) = \\ -\frac{N}{2}log(|\hat{\Sigma}_{diag}|) - Dlog(N) - \frac{N}{2}D \end{aligned} \tag{4}$$

Conclusion

In this exercise, we made a comparasion between the full covariance MVN and the diagonal covariance MVN using the BIC criteria for model selection. We note two differences between the BIC scores. First, the BIC penalize the complex model with a quadratic function on the number of dimensions D. Meanwhile, the simple model is penalized with a linear function on D. This is in accordance with our expectations and tells us that the BIC did in fact penalize the addition of new free parameters.

Another difference is in the $-\frac{N}{2}(log|\hat{\Sigma}|)$ term. Since this terms comes from the maximum likelihood value and it's the only difference between the ML of the two values, we can suspect that $-\frac{N}{2}(log|\hat{\Sigma}_{full}|)$ will be a greater number than $-\frac{N}{2}(log|\hat{\Sigma}_{diag}|)$, because it has more parameters to adjust to the data. But to be honest, I am not 100% sure if this holds true. If someone have something to say about it, please inform it in a new issue. I will love it to discuss the problem and learn more about it :).