Model selection

Gaussian model selection in the simplified Georgopoulos setting.

In this setting, we measure n times the same cell but each time with a different angle of movement $u_i = 2\pi(i/n)$ for $i = 0...(n-1)$. We decompose the regression function on a Fourier basis until size p with $2 * p + 1 \leq n$.

1. Create a matrix X of size n times $2 * p + 1$. For $u_i = 2\pi(i/n)$, the coefficient $X_{i,j}$ is given as follows
   - $X_{i+1,j} = 1$ if $j = 1$,
   - $X_{i+1,j} = \cos(ku_i)$ if $j = 2 * k$,
   - $X_{i+1,j} = \sin(ku_i)$ if $j = 2 * k + 1$.

2. By computing the different scalar products (with R), show that the columns of X are orthogonal but not of norm 1 and renormalize them: this gives you the matrix X'.

3. Give, for a given $d < p$, the projection estimator of the regression function composed of the first $2 * d + 1$ Fourier coefficients. Transform this into a function in R.

4. Simulate two different experiments:

$$Y_i = 16 + 14\cos(u_i) + 5\varepsilon_i \quad \text{and} \quad Y_i = 10 * \exp(-(u_i - \pi)^2/0.2) + 1 * \varepsilon_i$$

   with ε_i's i.i.d $N(0,1)$. Plot the data, the true function to estimate and 4 or 5 different projection estimators in each cases. Explain the problem of overfitting and the problem of taking a model of too low dimension. *NB: you can try other regression function if you want*

5. Make a function in R which, for a given p, computes the Mallow's Cp criterion for all the models ($d \leq p$) and gives the estimator which minimizes the criterion. *NB : the behavior would be similar for other models with - log-likelihood and AIC criterion*

6. Let us now fix p and look at all the subspaces V that can be written on the basis up to $2*p+1$. For instance, we could have a subspace generated by $1, \cos(u), \sin(2u)$. Let us look at the BIC criterion and assume σ^2 is known. Simplify the formulas to show that this minimization problem is solved by taking as non-zeros coordinates the ones for which $| < Y|e_i > |$ is larger than $\sqrt{\ln(n)\sigma^2}$. Implement this method and show the resulting estimator on both previous cases.