PHYS-GA2000-PS5

Ahmet Koral Aykın

October 17, 2023

1 Introduction

In this problem set, we are mainly interested in linear regression using singular value decomposition (SVD). SVD is a form of matrix decomposition in the following way

$$A = UWV^T \tag{1}$$

where U and V^T are orthonormal matrices while W is a digonal one. Once we get the matrices U, W, and V^T, it is trivial to find the inverse of the matrix A, which is given by

$$A^{-1} = VW^{-1}U^T \tag{2}$$

In question 2, the aim is to perform linear regression using SVD. And in question 1, we are asked to find specific values of the gamma function, $\Gamma(a)$.

2 Methods

The method used for the perform integration in question 1 is Gaussian quadrature. Detailed explanation of it can be found in the report of the previous PS.

In question 2, to get the inverse of the design matrix, A, we use SVD as stated earlier.

3 Results

The integrand in question 1 is given by

$$f(x) = x^{a-1}e^{-x} \tag{3}$$

The plot of the integrand for $a = 2, 3, 4$ are presented in Figure 1.

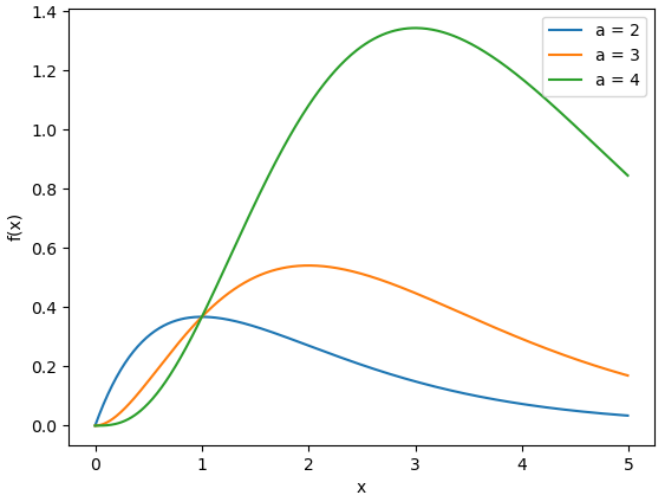

Figure 1: $f(x)$ for $a = 2, 3, 4$.

To find where the peak falls at, we need to set the first derivative of the integrand to 0

$$f'(x) = x^{a-2}e^{-x}(a - 1 - x) = 0 \tag{4}$$

whose solution is $x = a - 1$. It is clear from Fig.1 that this extremum is a maximum. The change of variables used has the following form

$$z = \frac{x}{x + c} \tag{5}$$

If we solve the above expression for c

$$c = \frac{x - xz}{z} \tag{6}$$

If we plug $z = 1/2$ and $x = a - 1$ such that the peak now is in the middle of the integration domain which is from 0 to 1 as a result of the change of variables, we can find $c = a - 1$. Also, it is of great importance to write the integrand as a single exponential since in this way the possibility of x^{a-1} being too big or e^{-x} too small are eliminated (a typical cause of round-off error). To put it another way, the terms in the exponent now have similar order of magnitude. The integrand can be written as a single exponential expression as follows

$$f(x) = e^{-x+(a-1)ln(x)} \tag{7}$$

The values of $\Gamma(a)$ for $a = 3/2, 3, 6, 10$ are calculated as 0.8862694302378622, 2.00000226e+00, 1.19999975e+02, 3.62880233e+05, respectively.

The plot of the data is presented in Figure 2.

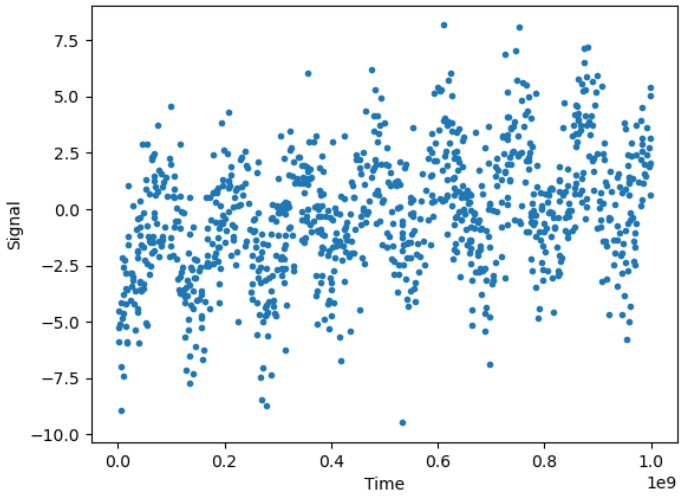

Figure 2: Signal vs. Time.

The third order polynomial fit and corresponding residual plot are presented in Figure 3 and 4, respectively. This is not a good fit to data. The uncertainty of every measurement is given in the problem statement as 2. As seen, the residuals are way above the associated uncertainty. Moreover, they still contain some information about the characteristic of the data.

It should be noted that the time axis is normalized by dividing the max value. Otherwise, the condition number of the design matrix will become too big such that it becomes unstable. There are different ways of doing this. However, as long as the values are near the unity, any rescaling should work.

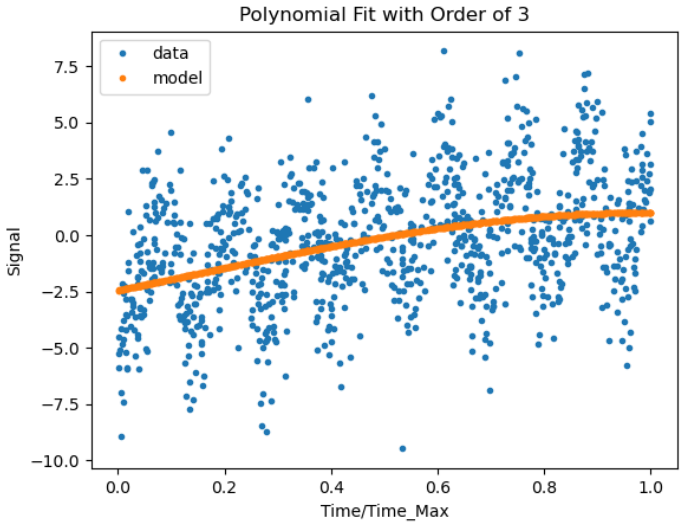

Figure 3: Signal vs. Time. Blue points corresponds to actual data points while orange is used for the polynomial fit with order of 3.

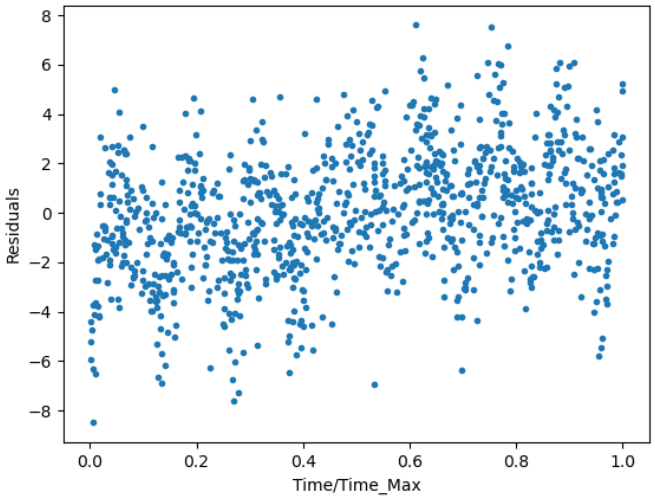

Figure 4: Residuals.

The polynomial fit of much bigger order (30th order) is presented in Figure 5. It can be suggested that polynomial fit to data which has a periodic behaviour is

not reasonable because of multiple reasons. First of all, polynomial expressions are unbounded. Also, periodic signals usually (of course it depends on the shape of the signal) have infinitely many extremum (like sin and cos) whereas an nth order polynomial has only n - 1. So, one needs to deal with really high order of polynomials to fit a periodic signal as in our case. In addition, the condition number of the design matrix for 30th order polynomial is on the order of 10^16 and it increases very fast with increasing order, which is not an ideal behavior.

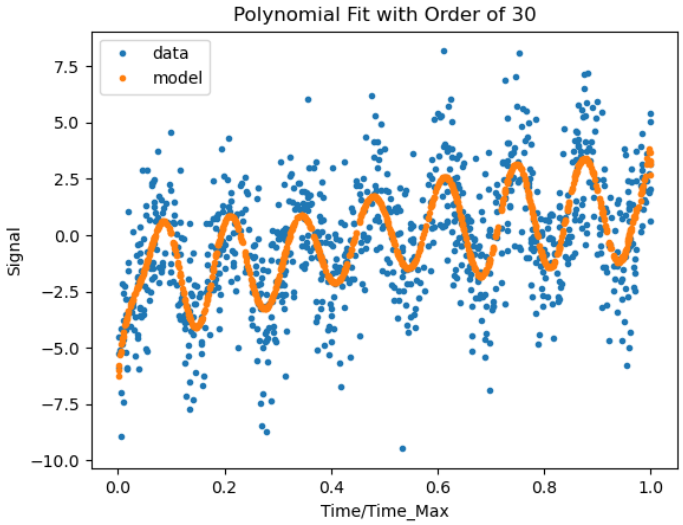

Figure 5: Signal vs. Time. Blue points corresponds to actual data points while orange is used for the polynomial fit with order of 30.

The cos and sin set plus zero point offset fit to data is presented in Figure 6. The condition number is on the order of unity which is very good, especially compared to polynomial case. The dominant period in the data revealed as result of the fit is equal to $T/8$ where T is the time span covered. This implies 8 repetition and it can be visually confirmed.

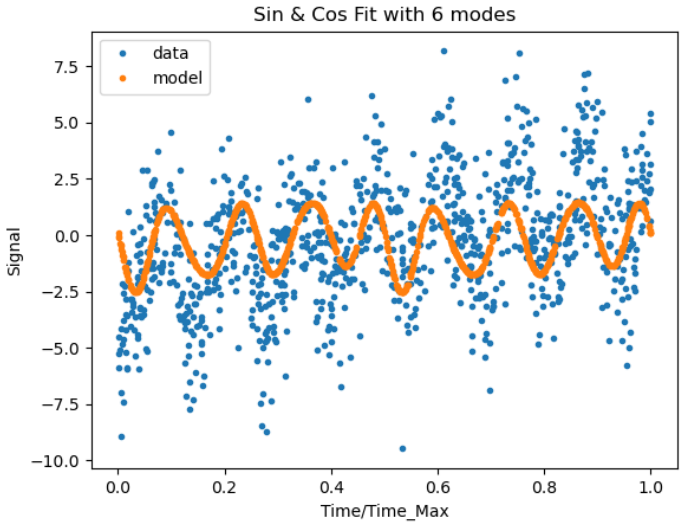

Figure 6: Signal vs. Time. Blue points corresponds to actual data points while orange is used for cos-sin set plus zero point offset fit with order of 30.