Introduction to Machine Learning - Lab 6

Gustav Sternelöv and Vuong Tran

30 november 2015

Assignment 1

(a)

The implementation of a function that simulates from the posterior distribution *f(x)* by using the squared exponential kernel is done in two steps. In the first step a function that computes the squared exponential kernel is created. The formula for the squared exponential kernel can be seen below:

$$K(x, x') = \sigma_f^2 \times exp(-0.5 \times (\frac{x - x'}{\iota})^2)$$

The second step is to build the function *PosteriorGP*. The aim with this function is to calculate the posterior mean and variance of f over a grid of x-values. The two formulas used for calculating this are presented below:

$$\bar{f}_* = K(x_*, x)[K(x, x) + \sigma^2 l]^{-1} y$$
$$cov(\bar{f}_*) = K(x_*, x_*) - K(x_*, x)[K(x, x) + \sigma^2 l]^{-1} K(x, x_*)$$

The code that has been used to implement the functions can be seen in the appendix *R-code*.

The prior mean of f is assumed to be zero for all x, which gives the following prior distribution for *f(x)*:

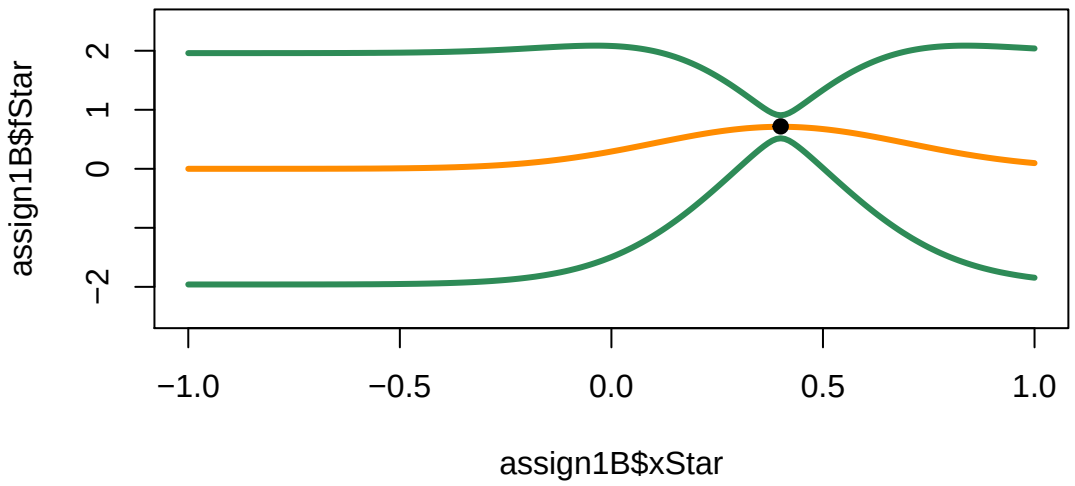
$$f(x) \sim GP(0, K(x, x'))$$

Then, the posterior guassian distribution looks as following:

$$f_* \mid x, y, x_* \sim N(\bar{f}_*, cov(\bar{f}_*)$$

(b)

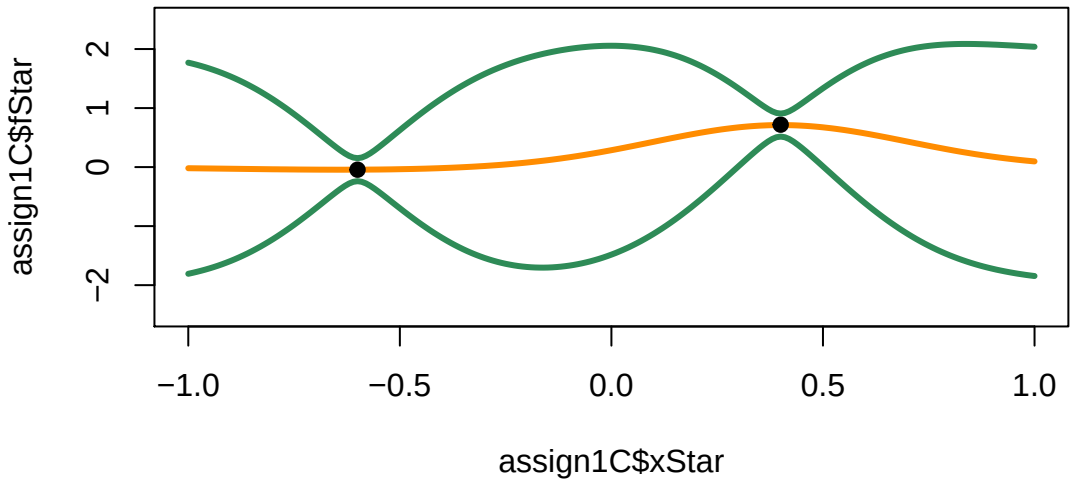
Since the noise standard deviation is assumed to be known the parameter σ_n is set to 0.1. The prior hyperparameter σ_f is set to 1 and the second prior hyperparameter ι is set to 0.3. Furthermore the prior is updated with one observation, *(x,y)*=(0.4, 0.719). A plot over the posterior mean of f over the interval $x \in$ [-1,1] with 95 % probability bands for f can be seen below.



We can see that the points narrow down the possibility values for where the true point could be, this is very much expected since more data leads to more estimated precision.

(c)

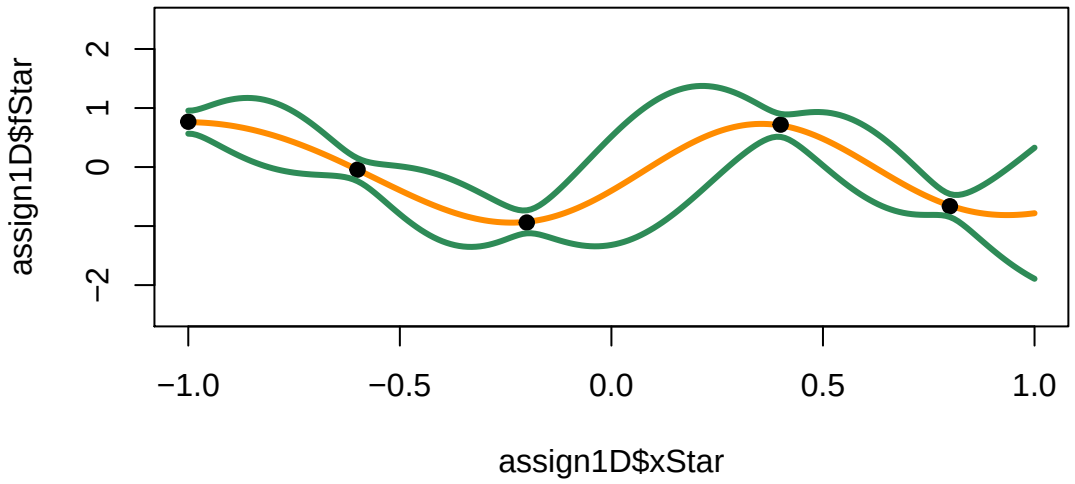
The posterior from *b)* is updated with another observation, *(x,y)*=(-0.6, -0.044).



Again it can be seen that the probability bands are more narrow around the observed values, and that they are quite wide for the other values.

(d)

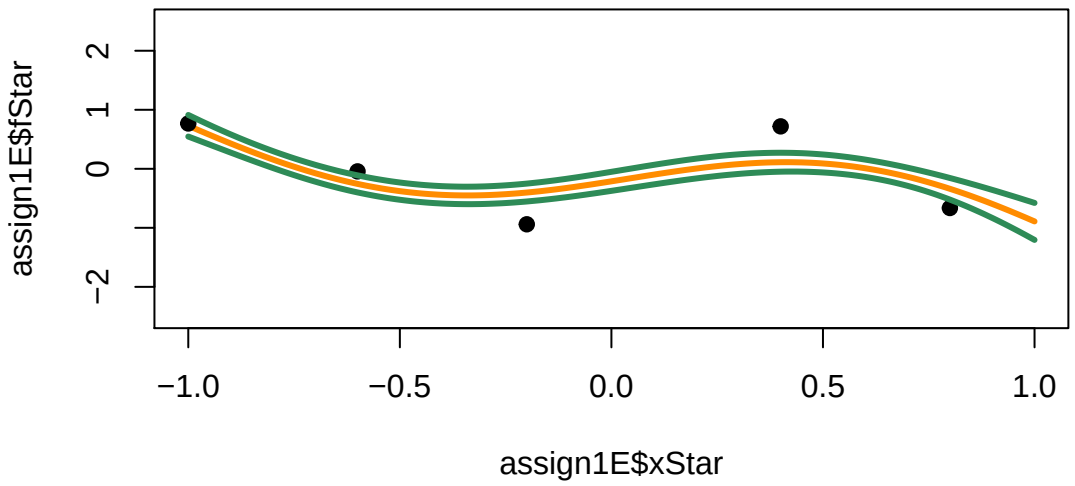
In *d)* the number of observations rises to five, resulting in the following plot over the posterior mean of f and its 95 % probability intervals.



Compared to the plots in *b)* and *c)*, the curve for the posterior mean of f is less straight/ more curvaceous than before. The probability bands has also changed and are thanks to the rise from two to five observed values more narrow, but also quite wiggly.

(e)

The hyperparameter ι is now set to 1. The other parameters are unchanged and the same observations as in *d)* are used.

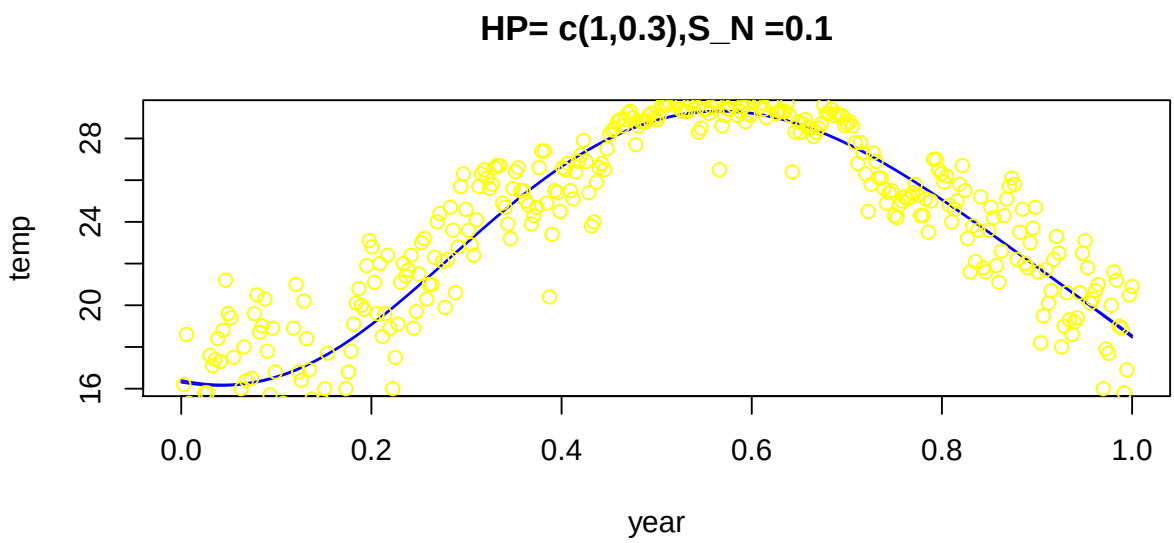


As we can see in the plot above, letting the length scale value increase to one make our fitted posterior mean really smooth and narrows down the 95 % probability bands quite much, maybe even to much since now we have three true observations that are outside the probability bands.

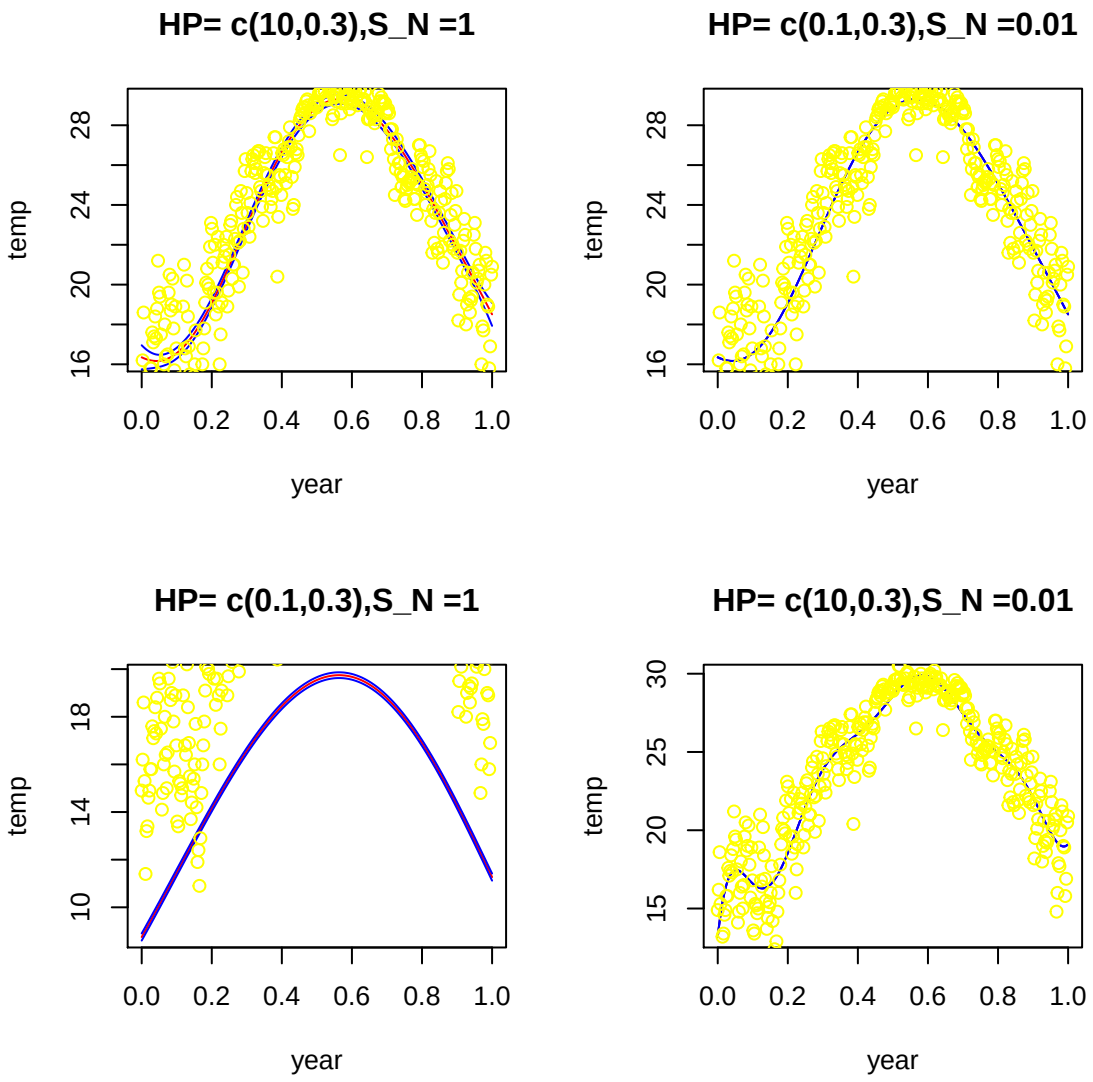
Assignment 2

The implemented functions in assignment 1 are now tested on the data set *JapanTemp*. This data set contains information about the daily temperatures during a year for some place in Japan. What that is mainly investigated in this assignment is the effect on the posterior for different values of the parameters σ_n, σ_f, and ι.

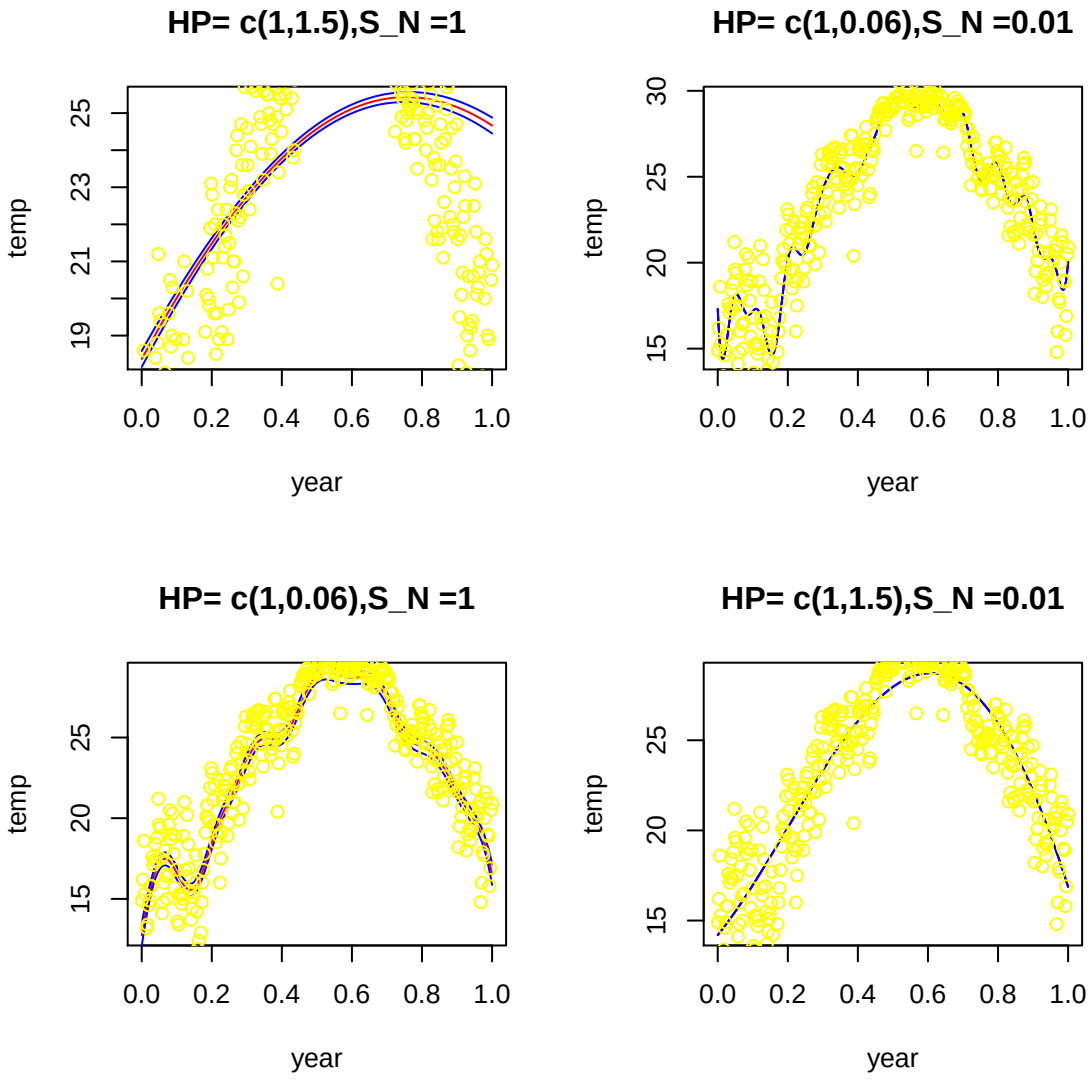
We started with plotting the data by using our GP-function with the prior hyper parameters $\{\sigma_f$=1, ι=0.3$\}$ and σ_n= 0.1 to extract the fitted point estimation. (this will be the "original" plot that we will make our comparisons against later on)



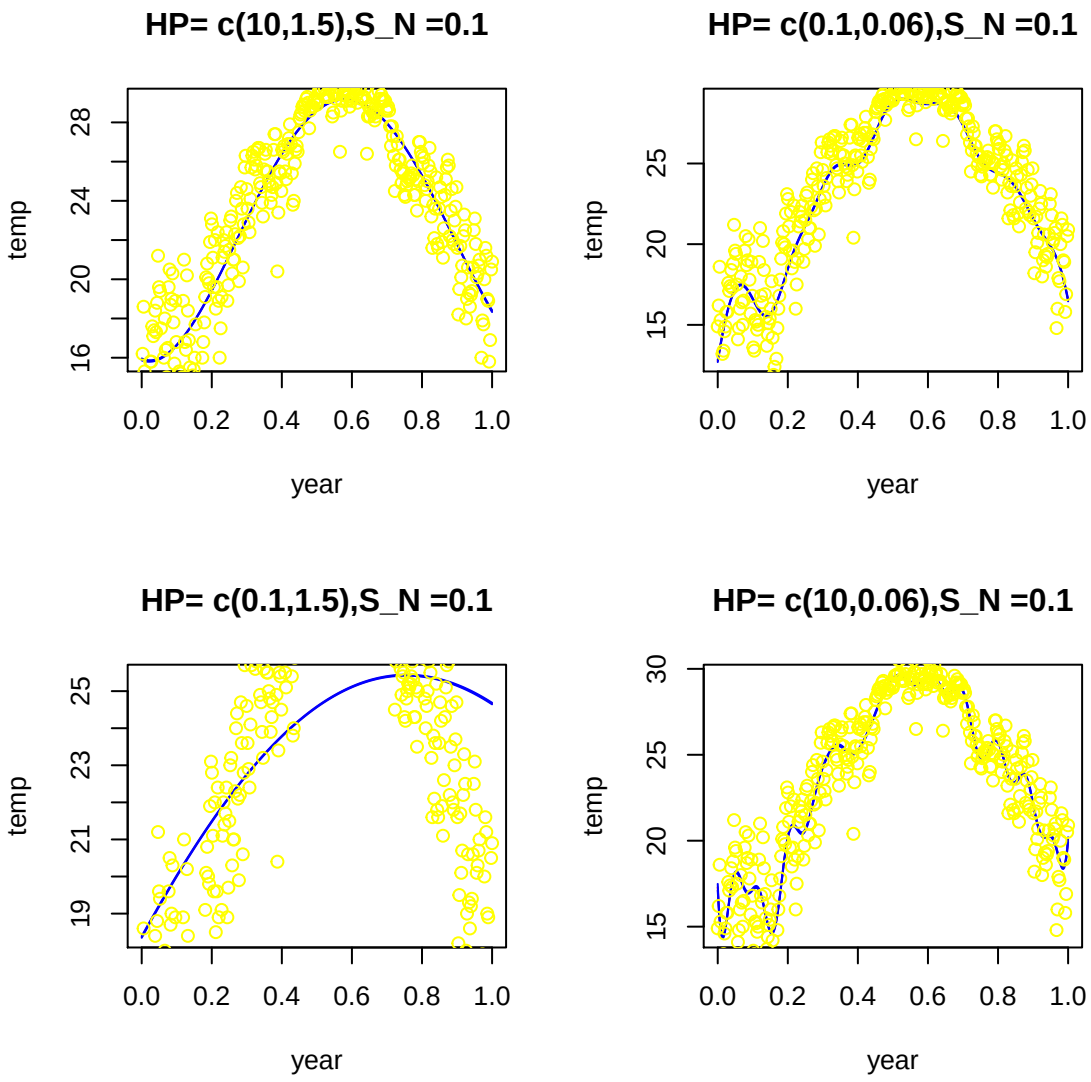
As one can see at the plot above, the fitted line and its 95 % probability band adapt really well for this data. Now, we will plot new fitted line and probability bands with different value for the hyper parameters and σ_n. The changes will have this structure, one of the parameter will be held as a constant while the rest of them first increased at same time, then decreased at the same time, and after that one parameter will decrease while the other parameter increase and at last the reverse.



We started out with letting the length scale (ι) being a constant equal to 0.3 and variating both the sigmas. The influence that σ_n have is easy to see, when increasing the σ_n parameter we get a wider probability band and while decreasing it we get narrower band. The effect of changing the value of σ_f looks quite similar to the effect of changing the value of σ_n with the difference that the probability bands remains practically unchanged. Also higher values of σ_f seem to give better fits with more flexible curves, compared to σ_n where lower values resulted in better fits.



As we hold the σ_f as a constant equal to one and variating the rest we can see that the length scale change the smoothness of the fitted curve. Increasing the length scale value leads to smoother lines, decreasing length scale value leads to more jagged curves. As like before, when increasing the σ_n parameter we get a wider probability band and while decreasing it we get narrower band.



The plots above have been plotted by letting the parameter σ_n be a constant equal to 0.1 and variating the rest. The appearance of the plot above is very similar to the graph we have plotted earlier. The influence of σ_f and the parameter length scaled is quite easy to as they look just like earlier findings.

Conclusion based on the given plots: The influence that σ_f has is that lower values gives more smooth curves and higher values results in more flexible curves. The length scale parameter have great influence on the smoothness of the fitted curve while σ_n have great influence on the width of 95 % probability band for the curve. If analyzing the plots more in depth, one could say that HP=$\{\sigma_f$=1,ι=0.06$\}$,σ_n=0.01 have the same influence as HP=$\{\sigma_f$=10,ι=0.06$\}$,σ_n=0.1. This is implying that holding σ_f constant and decreasing σ_n k times has same effect as holding σ_n as costant and increasing σ_f k times.