Machine Learning in Computational Biology
Master of Science – "Data Science and Information Technologies"
Academic year 2020-2021
Prof. Elias Manolakos

Supervised Learning
Proofs & Notes on Statistical Machine Learning

Aspasia Vozi

Contents

1 Maximum Likelihood Estimate ($\boldsymbol{\mu}_{ML}$)

Let $x_1, \ldots, x_n$ be N random vectors following a multidimensional Normal distribution. Assuming that the covariance matrix is known derive analytically the Maximum Likelihood Estimate "$\boldsymbol{\mu}_{ML}$" for the distribution's mean.

Using Maximum Likelihood method, and assuming that the covariance matrix $\boldsymbol{\Sigma}$ of $X = (x_1, \ldots, x_n)$ is known, we can estimate the distribution's mean.

1. First, we calculate the logarithm of the likelihood function:

$$
\begin{aligned}
\ln p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \ln\Big(\prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_{|}\boldsymbol{\mu}, \boldsymbol{\Sigma})\Big) \\
&= \ln\Big(\prod_{n=1}^{N} \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^1/2} exp\Big\{-\frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})\Big\}\Big) \\
&= \sum_{n=1}^{N} \ln\Big(\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^1/2} exp\Big\{-\frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})\Big) \\
&= \sum_{n=1}^{N} \ln\Big(\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^1/2}\Big) + \sum_{n=1}^{N} \ln exp\Big\{-\frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})\Big\} \\
&= N\ln\Big(\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^1/2}\Big) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu}) \\
&= -N\ln((2\pi)^{D/2}|\boldsymbol{\Sigma}|^1/2) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu}) \\
&= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})
\end{aligned}
$$

2. Second, we take the derivative of the log-likelihood over $\boldsymbol{\mu}$ and set it equal to 0, we obtain the solution for the maximum likelihood estimate of the mean, given by:

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 &\Leftrightarrow \frac{\partial}{\partial\boldsymbol{\mu}}\Big(-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})\Big) = 0 \\
&\Leftrightarrow \frac{\partial}{\partial\boldsymbol{\mu}}\Big(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu})\Big) = 0 \\
&\Leftrightarrow -\frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x_n}-\boldsymbol{\mu}) = 0 \\
&\Leftrightarrow \sum_{n=1}^{N}\mathbf{x_n} - N\boldsymbol{\mu} = 0 \\
&\Leftrightarrow \boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n}
\end{aligned}
$$

Evaluating the expectations of the maximum likelihood solutions under the true distribution, we obtain the following results:

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \frac{1}{N}\mathbb{E}\Big[\sum_{n=1}^{N}\mathbf{x_n}\Big] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathbf{x_n}] = \frac{1}{N}N\boldsymbol{\mu} = \boldsymbol{\mu}$$

Thus, similar to the univariate Gaussian distribution, the expectation of the maximum likelihood estimate for the mean is equal to the true mean.

2 Binomial distribution: Mean and Variance

Derive analytically the formulas for the mean and the variance of the Binomial distribution.

Let $X = x_1, \dots, x_N$ be N random variables that follow a Binomial Distribution $\mathcal{B}(N, \mu)$, where N is the number of trials and it is known, $\mu \in [0, 1]$ is the success probability of each trial and m is the number of successes. The probability of m is given by the following equation:

$$p(m, N, \mu) = Bin(m|N, \mu) = \binom{N}{\mu}\mu^m(1-\mu)^{N-m}$$

1. The average value of some function $f(x)$ under a probability distribution $p_x(x)$ is called the **expectation** of $f(x)$ and is denoted by $\mathbb{E}[f(x)]$. The average is weighted by the relative probabilities of the different values of x as follows:

   $$\mathbb{E}[f(x)] = \sum_x p_x(x)f(x) = \int p_x(x)f(x)dx$$

   We calculate the **mean** of the Binomial distribution using the equation above, as follows:

$$
\begin{aligned}
\mathbb{E}[m] &= \sum_{m=1}^{N} m\mathcal{B}(m|N,\mu) \\
&= \sum_{m=0}^{N} m\binom{N}{m}\mu^m(1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} m\binom{N}{m}\mu^m(1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} m\frac{N!}{(N-m)!m!}\mu^m(1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} m\frac{N(N-1)!}{[(N-1)-(m-1)]!m(m-1)!}\mu\mu^{m-1}(1-\mu)^{(N-1)-(m-1)} \\
&= \sum_{m=1}^{N} m\mu\frac{N}{m}\frac{(N-1)!}{[(N-1)-(m-1)]!(m-1)!}\mu^{m-1}(1-\mu)^{(N-1)-(m-1)} \\
&= N\mu\sum_{m=1}^{N}\binom{N-1}{m-1}\mu^{m-1}(1-\mu)^{(N-1)-(m-1)} = N\mu
\end{aligned}
$$

2. Whereas the expectation provides a measure of centrality, the **variance** of a random variable quantifies the spread of that random variable's distribution. Thus, the variance provides a measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$ and is defined as follows:

$$
\begin{aligned}
var[f(x)] &= \mathbb{E}[(f(x)-\mathbb{E}[f(x)])^2] \\
&= \mathbb{E}[(f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\
&= \mathbb{E}[(f(x)^2] - \mathbb{E}[2f(x)\mathbb{E}[f(x)]] + \mathbb{E}[\mathbb{E}[f(x)]^2] \\
&= \mathbb{E}[(f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\
&= \mathbb{E}[(f(x)^2] - \mathbb{E}[f(x)]^2
\end{aligned}
$$

Since *Expectation* has been calculated , $\mathbb{E}[m] = N\mu$, we calculate the *Variance*:

$$
\begin{aligned}
var[m] &= \mathbb{E}[m^2] - \mathbb{E}[m]^2 \\
&= \mathbb{E}[m^2] - (N\mu)^2 \qquad (1)
\end{aligned}
$$

We have to calculate $\mathbb{E}[m^2]$:

$$\begin{aligned}
\mathbb{E}[m^2] &= \sum_{m=0}^{N} m^2 \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} mm \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} mN \binom{N-1}{m-1} \mu\mu^{m-1} (1-\mu)^{(N-1)-(m-1)} \\
&= N\mu \sum_{m=1}^{N} m \binom{N-1}{m-1} \mu^{m-1} (1-\mu)^{(N-1)-(m-1)} \\
(N-1 = a \;\&\; m-1 = b) &= N\mu \sum_{b=0}^{a} (b+1) \binom{a}{b} \mu^b (1-\mu)^{a-b} \\
&= N\mu \Big[\sum_{b=0}^{a} b \binom{a}{b} \mu^b (1-\mu)^{a-b} + N\mu \sum_{b=0}^{a} \binom{a}{b} \mu^b (1-\mu)^{a-b} \Big] \\
&= N\mu(a\mu + 1) = N\mu[(N-1)\mu + 1] \\
&= N\mu(N\mu - \mu + 1) = (N\mu)^2 + N\mu(1-\mu)
\end{aligned}$$

We can now calculate $var(m)$ from equation (1):

$$\begin{aligned}
var[m] &= \mathbb{E}[m^2] - (N\mu)^2 \\
&= (N\mu)^2 + N\mu(1-\mu) - (N\mu)^2 \\
&= N\mu(1-\mu)
\end{aligned}$$

3 Posterior Distribution Estimation

Let x be a random variable following a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with a known variance σ^2 but an unknown mean μ. As Bayesians, we believe that μ follows a prior distribution $\mathcal{N}(\mu_0, \sigma_0^2)$. Provided that we are given a data set of N independent observation, $X = x_1, \ldots, x_n$ show that:

Exercise 3.1

The posterior distribution of the mean $p(\mu|X)$ is also a Gaussian with mean $\mu_N = \frac{N\sigma_0^2\overline{x} + \sigma^2\mu}{N\sigma_0^2 + \sigma^2}$, where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, and variance $\sigma^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$.

The likelihood of the distribution is given by:

$$p(\mathbf{x}|\mu, \sigma^2) = \Big(\prod_{n=1}^{N} \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\sigma^2}) \Big) = \frac{\mathbf{1}}{(\mathbf{2}\pi\sigma)^{\frac{\mathbf{N}}{\mathbf{2}}}} \mathbf{exp}\Big\{ - \frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}} \sum_{\mathbf{n=1}}^{\mathbf{N}} (\mathbf{x_n} - \mu)^{\mathbf{2}} \Big\}$$

The prior distribution is given by:

$$p(\mu) = \mathcal{N}\big(\mu|\mu_0, \sigma_0^2\big)$$

and the posterior distribution is given by:

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$$

The posterior distribution is proportional to the product of the likelihood and the prior. Due to the choice of a conjugate Gaussian prior distribution, the posterior is also Gaussian. Thus, to derive the form of the posterior, we focus on the exponential term:

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 \tag{2}$$

If we open all the quadratic forms, we have:

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 = -\frac{\sum_{n=1}^{N}x_n^2}{2\sigma^2} + \frac{N\mu\sum_{n=1}^{N}x_n}{\sigma^2} - \frac{N\mu^2}{2\sigma^2}$$

$$-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 = -\frac{\mu^2}{2\sigma_0^2} + \frac{\mu\mu_0}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2}$$

We can now write equation (2) as follows:

$$-\Big(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2}\Big)\mu^2 + \Big(\frac{N\boldsymbol{x}}{\sigma^2} + \frac{\mu_0}{2\sigma_0^2}\Big)\mu - \Big(\frac{\sum_{n=1}^{N}x_n}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\Big)$$

By definition we can equate the derived exponent to the exponent of the posterior, in this we way we have:

$$-\Big(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2}\Big)\mu^2 + \Big(\frac{N\boldsymbol{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\Big)\mu - \Big(\frac{\sum_{n=1}^{N}x_n}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2}\Big) = -\frac{1}{2\sigma_N^2}(\mu^2 + \mu_N^2 - 2\mu\mu_N)$$

Therefore, matching the coefficients of μ^2, we have:

$$\begin{aligned}
-\frac{1}{2\sigma_N^2}\mu^2 &= -\Big(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2}\Big)\mu^2 \Leftrightarrow \\
\frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \Leftrightarrow \\
\frac{1}{\sigma_N^2} &= \frac{\sigma_0^2 N + \sigma^2}{\sigma^2\sigma_0^2} \Leftrightarrow \\
\sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{\sigma_0^2 N + \sigma^2}
\end{aligned}$$

And matching coefficients of μ, we have:

$$
\begin{aligned}
\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} &= \frac{\mu_N}{\sigma_N^2} \Leftrightarrow \\
\frac{N\bar{x}\sigma_0^2}{\sigma_0^2\sigma^2} + \frac{\sigma\mu_0}{\sigma^2\sigma_0^2} &= \frac{\sigma_0^2 N + \sigma^2}{\sigma^2\sigma_0^2}\mu_N \Leftrightarrow \\
\frac{N\sigma_0^2\bar{x} + \sigma\mu_0}{\sigma^2\sigma_0^2} &= \frac{\sigma_0^2 N + \sigma^2}{\sigma^2\sigma_0^2}\mu_N \Leftrightarrow \\
\mu_N &= \frac{N\sigma_0^2\bar{x} + \sigma\mu_0}{\sigma_0^2 N + \sigma^2}
\end{aligned}
$$

We have obtained the formulas for the calculation of the posterior distribution, $\mu_N = \frac{N\sigma_0^2\bar{x}+\sigma\mu_0}{\sigma_0^2 N+\sigma^2}$, $\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 N+\sigma^2}$ and we are going to use them in the next tasks of the exercise 3.

Exercise 3.2

Consider now that x is distributed as $x \sim \mathcal{N}(\mu, 16)$, and as Bayesians, we believe that the prior for the mean is $\mu \sim \mathcal{N}(0, 4)$. Use the distribution $\mathcal{N}(7, 16)$ to generate observations for x.

Exercise 3.2.1

Develop an algorithm that estimates the posterior distribution's $p(\mu|X)$ mean and variance, assuming we have available $N = 1, 5, 10, 20, 50, 100$ and 1000 observations, respectively. What do you observe as the number of observations N is increasing?

In order to implement the algorithm, we used the equations of Exercise 3.1. to calculate the posterior distribution of the mean.
From Table 1, we note that as the number of observation increases, the posterior mean tends to became equal to its true value (μ_{true}=7). Moreover, the variance decreases as the number of observation increases, meaning our model becomes more ”certain” for its prediction. The aforementioned observation were expected.

Number of observations	μ_N	σ_N
1	1.34	3.2
5	4.13	1.77
10	5.419	1.14
20	6.93	0.66
50	6.36	0.29
100	5.83	0.15
1000	6.82	0.01

Table 1: Posterior distribution's **mean** and posterior distribution's **variance** and increasing number of observations.

Prior, Posterior & the Original Distribution

For every N, provide a diagram that shows the prior distribution, the distribution generating the data, and the estimated posterior distribution.

We can confirm from the plots, presented in Figure 1 below, that as the number of observation increases, the estimated posterior distribution approaches the mean of the distribution that generated the data as expected. When we try to approximate the distribution of the mean using only 1 or 5 sample data, the posterior distribution's position depends more on the already known prior distribution, and thus it is closer to it. As we increase the number of observations the posterior distribution depends more on the data provided.

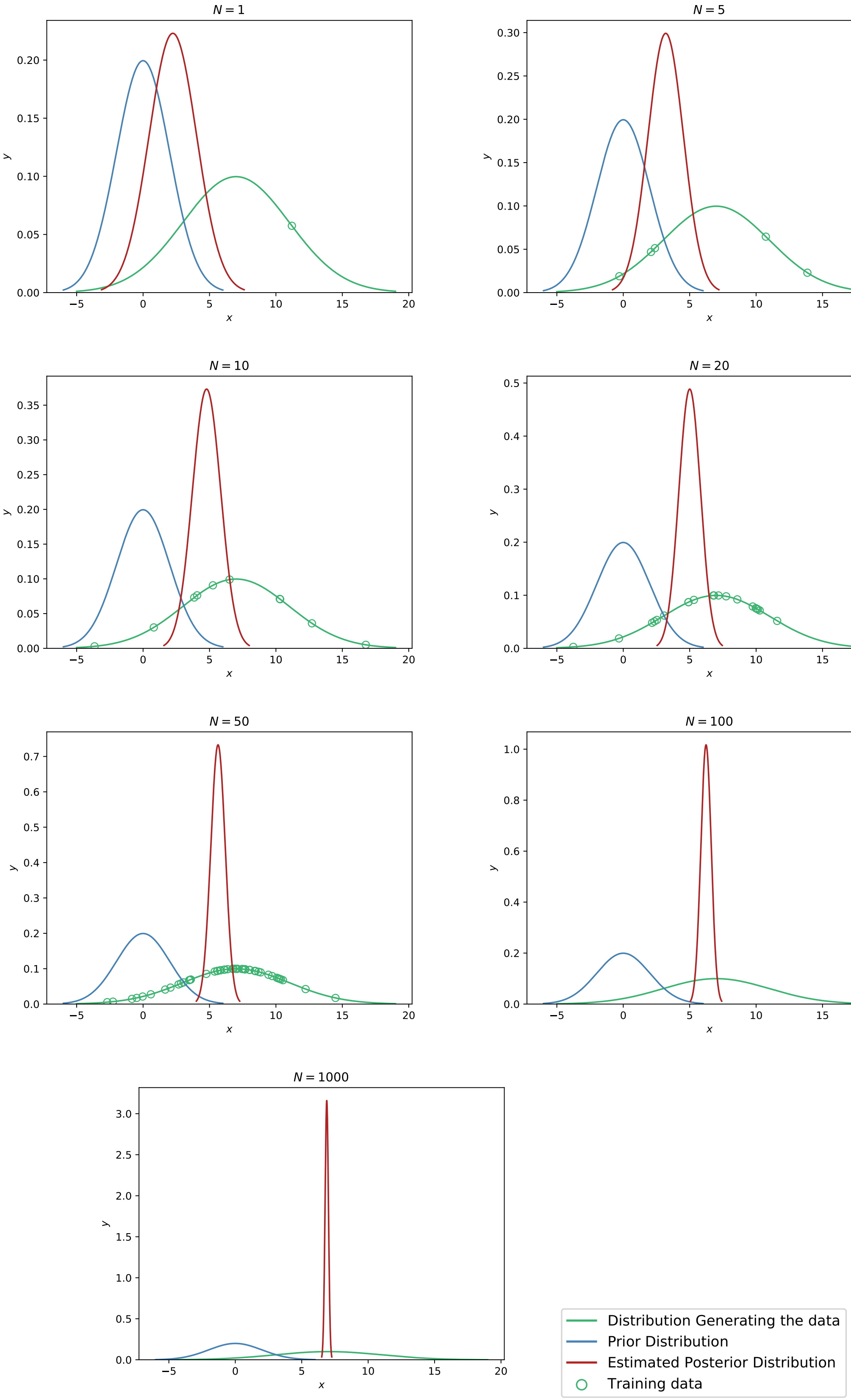

Figure 1: Prior Distribution (*blue*), the distribution generating the data (*green*) and the estimated posterior distribution (*red*) for $N = 1, 5, 10, 20, 50, 100, 1000$ observations.

4 The *Curve Fitting* Problem

Draw a period of the sinusoidal function $y(x) = \sin(2\pi x)$ and select N samples for x uniformly distributed in the interval $[0, 1]$. To every $y(x)$ value add Gaussian noise distributed as $\mathcal{N}(0, 1)$ to generate a set of observations.

Fit to the noisy observations a polynomial (in the data) model of degree $M = 2, 3, 4, 5$ or 9 and provide a table with the coefficients of the best least-squares fit model and the achieved RMSE. Also, provide a plot showing the function $y(x)$, the observations drawn, and the best fit model for every different value of M.

Repeat the above procedure for $N = 10$ and $N = 100$. What do you observe? Discuss your findings.

Theoretical Background

Our goal is to predict the value of $\hat{t}$ for some new value of $\hat{x}$, in the absence of any knowledge for the function that generates the data. To that end, we consider a simple approach based on curve fitting. In particular, we shall try to fit the data using a polynomial function of the form:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

where M is the degree of the polynomial and it is called linear model.
Next, we need to determine the values of the coefficients $\mathbf{w}$ by fitting the polynomial to the training data. This can be done by minimizing an error function that measures the misfit between the function $y(x, \mathbf{w})$, for a given value of $\mathbf{w}$, and the training data points. One simple error function is the sum of squares of the errors between $y(x, \mathbf{w})$ and the corresponding target values t_n

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(x, \mathbf{w}) - t_n)^2 \geq 0$$

, where the function becomes zero if, and only if, the function $y(x, \mathbf{w})$ were to pass exactly through each training data point.
We can solve the curve fitting problem by choosing the value of $\mathbf{w}$ for which $E(\mathbf{w})$ is as small as possible. Because the error function is quadratic, its derivatives are linear, and so the minimization of the function has a unique closed from solution, denoted by $\mathbf{w}^*$. To minimize the error function we should derive the gradient vector, set it equal to zero and solve for $\mathbf{w}^*$

$$\nabla E(\mathbf{w}^*) = \mathbf{0}$$

To that end, first, we have to substitute the polynomial into the error function

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^j - t_n)^2$$

To find the gradient vector, we take the partial derivative of E with respect to an arbitrary w_k. Differentiating the sum, term by term, we get:

$$\begin{aligned}
\nabla E(\mathbf{w}^*)_k &= \frac{\partial}{\partial w_k}(\mathbf{w}) \\
&= \frac{1}{2}\sum_{n=1}^{N} 2(\sum_{j=0}^{M} w_j x_n^j - t_n)x_n^k = \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)x_n^k \\
&= \sum_{n=1}^{N}(\mathbf{X}\mathbf{w} - \mathsf{t})_n \mathbf{X}_{nk} = \sum_{n=1}^{N} \mathbf{X}_{kn}^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathsf{t})_n \\
&= \left(\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathsf{t})\right)_k
\end{aligned}$$

Using the partial derivative for one component, we compute the gradient vector by dropping the k subscript. Thus, the minimizer $\mathbf{w}^*$ must satisfy:

$$\nabla E(\mathbf{w}^*) = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w}^* - \mathsf{t}) = \mathbf{0}$$

Solving for $\mathbf{w}^*$ gives the unique solution of the curve fitting problem

$$\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w}^* - \mathsf{t}) = \mathbf{0} \Leftrightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}^* = \mathbf{X}^{\mathrm{T}}\mathsf{t} \Leftrightarrow \mathbf{w}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathsf{t} \tag{3}$$

The resulting polynomial is given by the function $y(x, \mathbf{w}^*)$.

Results

We generate two training sets using $N = 10$ and $N = 100$ equidistant points in the interval $[0, 1]$. For every x, we use the true model:

$$y = \sin(2\pi x) \tag{4}$$

, we add i.i.d noise samples to every y(x) ,(η,) originating from a Gaussian distribution $\mathcal{N}(0, 1)$:

$$t = y(x) + \eta \tag{5}$$

For the purpose of the exercise we generate a set of 10 and 100 observations.

We fit our training data, for each of the two observations set, to equation (3) to obtain the Least squares estimates for $\mathbf{w}$ coefficients. In Table 2 are presented the different coefficients calculated for each polynomial and for the different number of observations: To measure the performance of our estimators we use the **Root Mean-Square Error**, which quantifies the deviation of the predicted output values given a test set from the expected values of t from the observations:

$$MSE = \mathbb{E}[(\hat{w} - w_0)^2]$$

We generate a test set of $T = 1000$ points in the interval $[0, 1]$. We calculate the predicted $\hat{t}$ values using the model function that we obtained before, for $x_i, i = 1, 2, \ldots, 1000$:

M	**N = 10**	**N = 100**
2	**w**=[0.02, 2.35, -3.51]	**w**=[0.78, -0.94, -0.67]
3	**w**=[-0.78, 15.70, -38.71, 23.46]	**w**=[-0.64, 16.64, -44.87, 29.46]
4	**w**=[-1.05, 26.09, -91.18, 107.61, -42.07]	**w**=[-0.34, 10.33, -16.11, -15.41, 22.43]
5	**w**=[-0.74, -2.27, 145.34, -560.97, 724.28, -306.54]	**w**=[6e-04, -7e-01, 6e+01, -2e+02, 2e+02,-9e+01]
9	**w**=[-7e-01, -2e+02, 6e+03, -6e+04, 2e+05, -7e+05, 1e+06, -1e+06, 6e+05 -1e+05]	**w**=[-2e-01, 2e+01, -5e+02, 4e+03, -2e+04, 5e+04, -8e+04, 8e+04, -4e+04, 1e+04]

Table 2: Predicted Coefficients, using Least Squares method for **different degrees of polynomial** $M = 2, 3, 4, 5, 9$ **and different sample sizes** $N = 10, 100$

For example, for the third degree polynomial, predicted using N=10 observations, we would use:

$$\hat{t} = -0.78 + 15.70x_i - 38.71x_i^2 + 23.46x_i^3$$

We also calculate the values of t from the true model (2) for every x_i in our test set, and then, we calculate the RMSE:

$$RMSE = \sqrt{\frac{1}{M}\sum_{t=1}^{1000}(\hat{t}_i - t_i)^2}$$

In Table 3 are presented the RMSE values for the different observation sets and the different degree of polynomial.

M	2	3	4	5	9
$N = 10$	12.50	4.98	4.25	6.99	16.30
$N = 100$	10.34	3.29	3.33	3.00	2.44

Table 3: Root Mean Square Error for **different degrees of polynomial** $M = 2, 3, 4, 5, 9$ **and different sample sizes** $N = 10, 100$**.**

We note that RMSE tends to decreases as the degree of polynomial increases when we have the bigger set of observations $N = 100$. Moreover, for the smaller set of observations $N = 10$, although initially RMSE decreases, as the degree of polynomial increases, we note the for $M = 5$ and $M = 9$, the RMSE increases dramatically. These observations can be better visualized in Figure 2.

In Figure 3 and Figure 4 are presented the plots, that show the function y(x), the observations used its time to train our model and the best fit model for the different values of M.

For the first case of $N = 10$ training points, we note that the second order (M=2) polynomial gives rather poor fit to the data. The third order (M=3) polynomial seems to give the best fit, while the higher order one (M=9) achieves an excellent fit to the data, that is, $E(\mathbf{w}^*)$=$\mathbf{0}$. However, the fitted curve gives a poor representation of the underlying function $\sin(2\pi x)$. The latter case is the outcome of the phenomenon that is known as over-fitting.

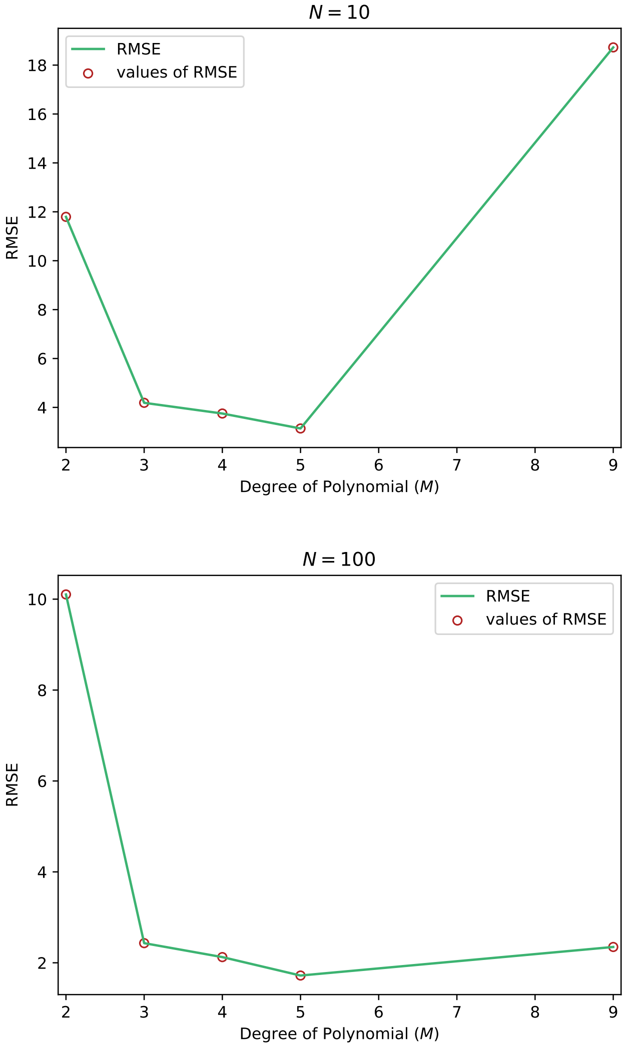

Figure 2: Root Mean Square Error for **different degrees of polynomial** $M = 2, 3, 4, 5, 9$ **and different sample sizes** $N = 10, 100$**.**

For the second case of $N = 100$ training points, we note that the second order (M=2) polynomial still gives poor fit to the data. However, not only the third order (M=3) polynomial, but also the fourth adn the fifth order polynomials seem to give good fit. In this case, the higher order one (M=9) achieves an rather good fit to the data, and the fitted curve gives a better representation of the underlying function $\sin(2\pi x)$ than in the first case.

We note that the over-fitting problem becomes less severe as the size of the training data set increases. In other words, the larger the data set, the more complex the model that we can afford to fit to the data. This can be also confirmed by observing the fitting curves of the three intermediate degree polynomials, M=3 ,M=4 and M=5 in the first small training set compared to the second bigger one.

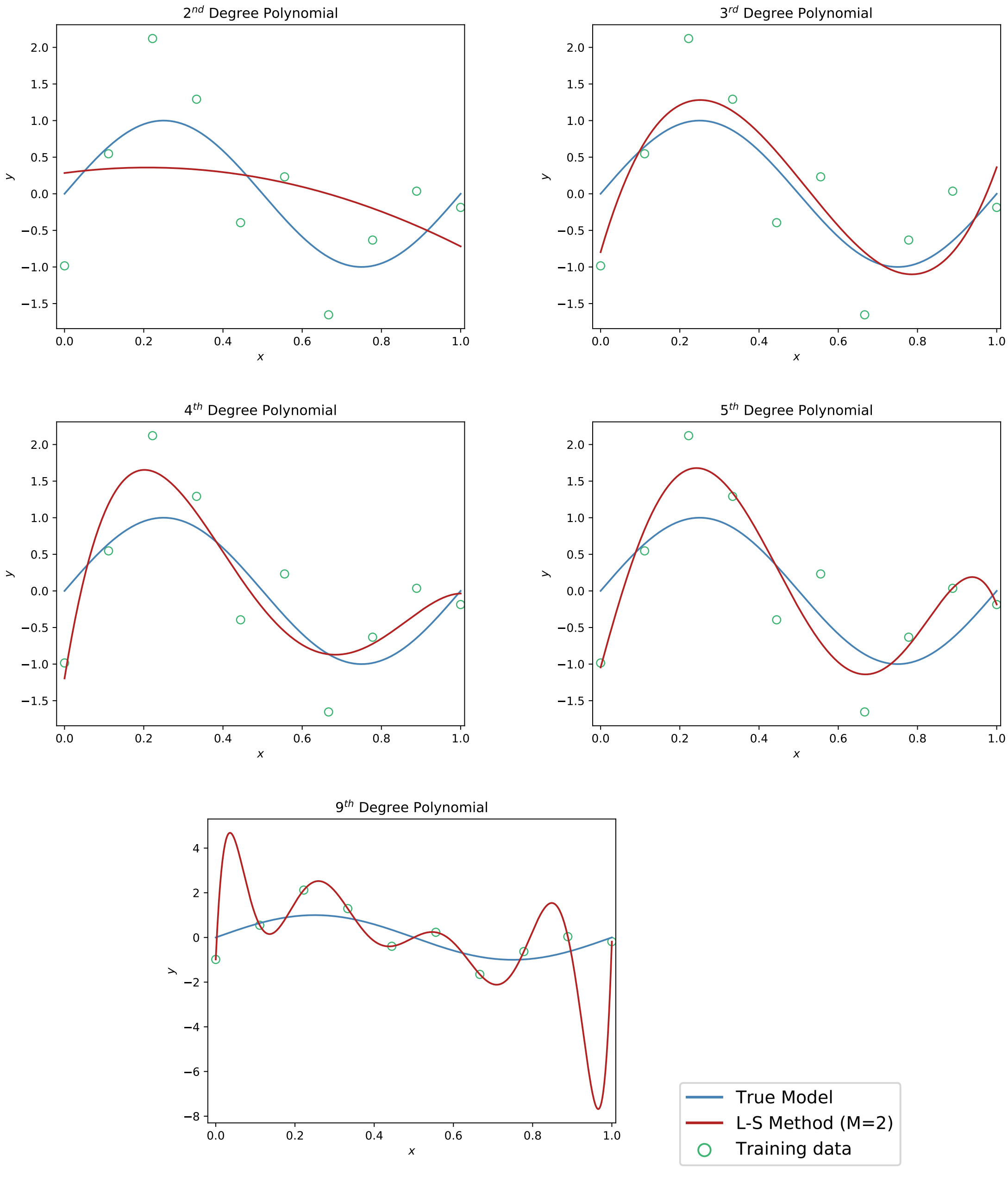

Figure 3: Least Squares Method fit model $y(x)$ for **different degrees of polynomial** $M = 2, 3, 4, 5, 9$, along with the true model, using $N = 10$ **training data**.

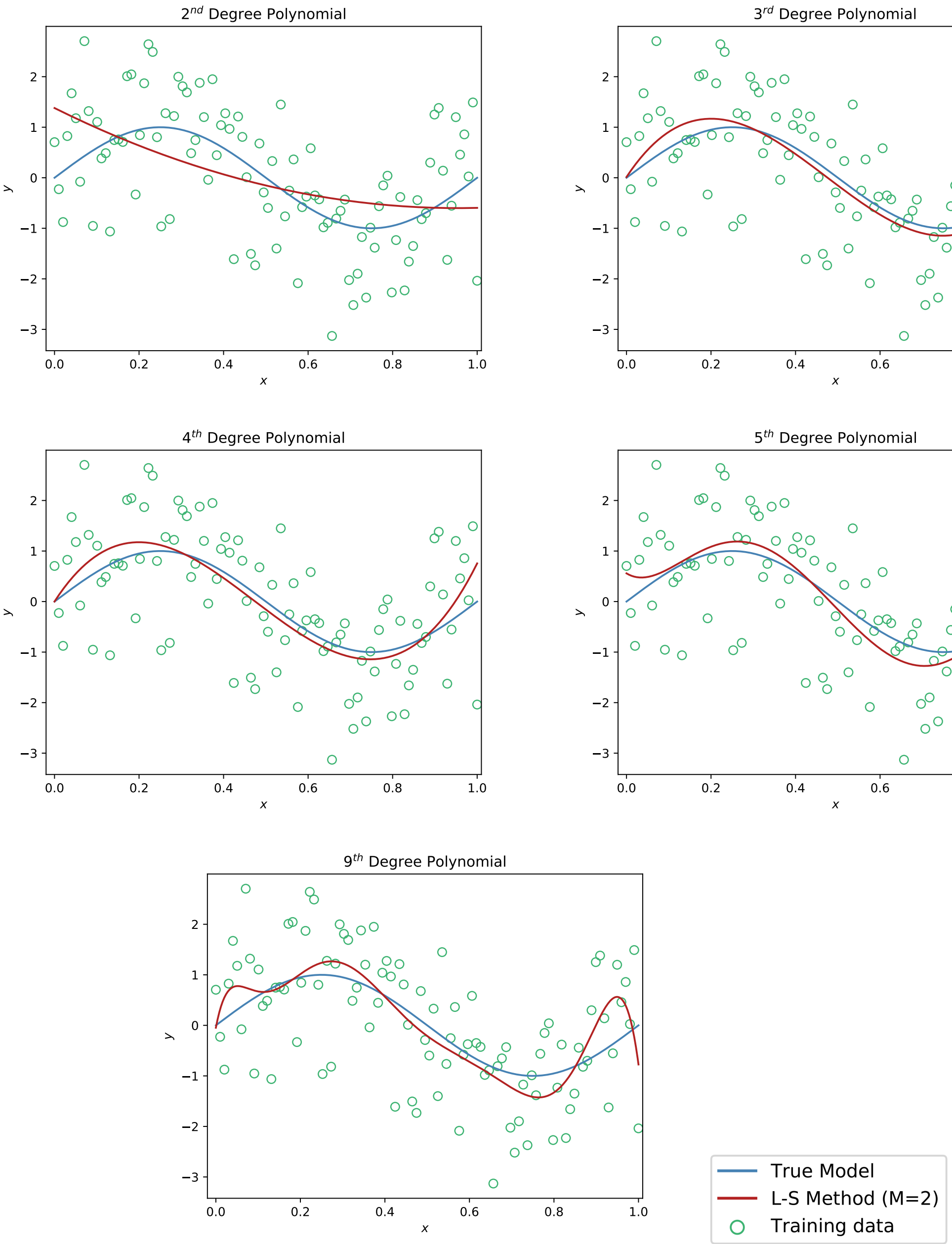

Figure 4: Least Squares Method fit model $y(x)$ for **different degrees of polynomial** $M = 2, 3, 4, 5, 9$, along with the true model, using $N = 100$ **training data**.

5 Gaussian Approach

For the same setup as in Problem 4 above, let 's assume that the observations are generated as $t = y(x) + \eta$, where $y(x) = \sin(2\pi x)$ and the Gaussian noise η is distributed by $\mathcal{N}(0, \beta - 1)$ with $\beta = 11.1$. You are given a dataset generated in this way with $N = 10$ samples (x, t) where $0 < x < 1$. Assume that you want to fit to the data a regression model of the form $t = g(x, w) + \eta$, where $g(x, w)$ is an $M = 9$ degree polynomial with coefficients vector w following a Normal prior distribution with precision $\alpha = 0.005$ (Bayes approach), i.e., the prior for w is

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) = \Big(\frac{\alpha}{2\pi}\Big)^{\frac{M+1}{2}} exp\Big\{-\frac{\alpha}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}\Big\}.$$

Construct the predictive model

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|\boldsymbol{\mu}(x), s^2(x))$$

which allows for e very unseen x (not in the training set) to produce a prediction t. Plot the mean $m(x)$ and variance $s^2(x)$ of the predictive Gaussian model for many different values of x in the interval $0 < x < 1$. What do you observe? Discuss your findings.

For the purpose of this exercise, we turn to the Bayesian treatment of linear regression that leads to automatic methods for determining model complexity using the training data alone.

Theoretical Background

Our goal is make predictions of t for unseen values of $\mathbf{x}_{\text{unseen}}$, and thus, we are not actually interested in the value of $\mathbf{w}$ itself. To that end, we evaluate the predictive distribution given by:

$$p(t|\mathbf{x}_{\text{unseen}}, \mathbf{t}, \mathbf{X}, \alpha, \beta) = \int p(t|\mathbf{x}_{\text{unseen}}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \mathrm{d}\mathbf{w}$$

We note that the predictive distribution involves the convolution of the conditional Gaussian distribution of the target variable and the posterior weight Gaussian distribution.

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) = \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \beta^{-1})$$
$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Taking advantage of the properties marginal and conditional Gaussians' properties, we can obtain:

$$p(t|\mathbf{x}_{\text{unseen}}, \mathbf{t}, \mathbf{X}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{\text{unseen}}), \sigma_N^2(\mathbf{x}_{\text{unseen}}))$$

, where the mean $m(\mathbf{x})$ and variance $s^2(\mathbf{x})$ of the predictive distribution is given by:

$$m(\mathbf{x}) = \beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x_n})t_n$$

$$s^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$$

, where

$$S^{-1} = a\boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x_n})\boldsymbol{\phi}(\mathbf{x_n})^{\mathrm{T}}$$

In the equation of the variance, the first term represents the noise on the data whereas the second term reflects the uncertainty associated with the parameters $\mathbf{w}$ and it is a consequence of the Bayesian treatment.

Results

In Figure 5, we plot the results after fitting our Gaussian model for a training set of N=10. For each value of a test set of T=1000 points, we plot the predicted $m(x)$ and variance $s^2(x)$, the true value of y(x), given by equation (2), and the training points used to train our model.

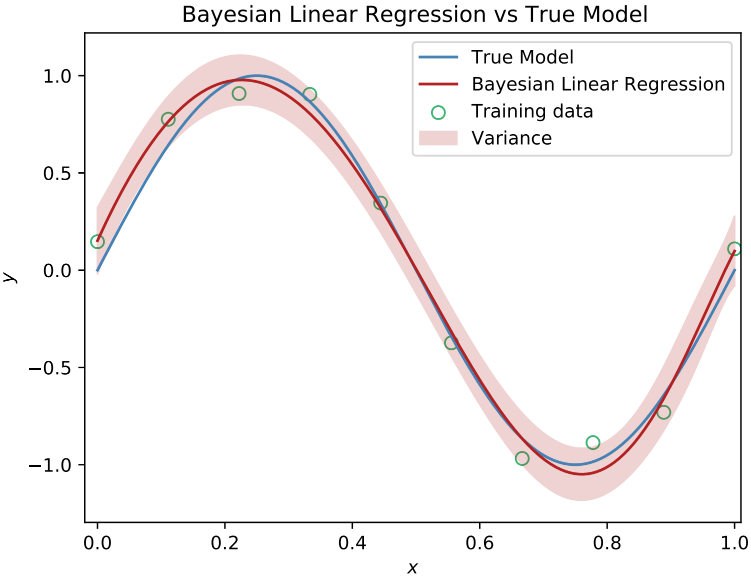

Figure 5: Predictions of $m(x)$ by the bayesian approach model and their variance, the values of the true model for T=1000 points and training data.

We note that for a small training set, our predictive model gives a very good fit to the data. It's predicted value are very close to the model. Moreover we observe that the noise, that we have added to our true model, is covered by our model. Since all the training points are inside the variance boarders.