Gaussian Process Regression

Rohit Jain

Mentor - Dr. Swanand Ravindra Khare

Report for B.Tech Project

Department of Mechanical Engineering
Indian Institute of Technology
India
10th November, 2017

Department of Mechanical Engineering

Indian Institute of Technology Kharagpur

Certificate

This is to certify that this is a bonafide record of the project presented by Rohit Jain(Roll no. 14MF3IM13) during Autumn 2017 in partial fulfillment of the requirements of the degree of Bachelor of Technology in Manufacturing Science And Engineering as a part of his Bachelor's Thesis.

Prof S.R.Khare
(Project Guide)

Prof. Anandroop Bhattacharya and Prof. Atul Jain
(Course Coordinator)

Date:10th November 2017

1 Abstract

In our daily life we come across many situations where we need to predict something so as to prevent a cause or optimize the outcome. This is where predictive modelling and data analysis comes into picture. In the starting sections of the study, basics of probability and related rules have been discussed. Then a firm base is made on the modelling and procedural part of the algorithm. Also, an important application of this algorithm is shown in determining different properties of diesel. Lastly a discussion over the universality of the algorithm, scope for improvement and future work is shown in the concluding part.

2 Introduction

There are mainly 2 types of predictive models, regression and classification. Regression deals with prediction of continuous functions and classification deals with discrete function prediction. For this there are many famous algorithms such as linear regression, naive Bayes, decision trees etc. that are used in most of the applications. Most of them are deterministic in nature. But these models needs to be trained or made to learn the pattern or trend that it follows. The main methods of training used for all of them being some kind of loss reduction(or error reduction) like in the case of linear regression, RMSE(root mean square error) optimization. But there are many algorithms which do not follow this type of optimization technique such as Gaussian process regression[RW06]. The approach used for the optimization for prior mentioned methods are derivative based(some form of differential form). But this is not the case with Gaussian process regression. It is a Bayesian based method as it uses prior and posterior probabilities for its prediction[WR06].

3 Literature Review

The books 'Introduction to mathematical statistics[HC95]' and 'An introduction to probability theory and its applications[Fel08]' provides lot of basic knowledge about the concepts of probability and statistics needed to understand each and every bit used in the related papers including the multivariable conditional probability[HC95] discussed here. The kernel based method

and basics about GPR is also derived from 'Gaussian processes for machine learning'[RW06] and 'Advances in Gaussian processes' [Ras06].
The main detailing about the NRI spectroscopy and its relation to this field could be seen from 'Rapid analysis of diesel fuel properties by near infrared reflectance spectra[FWZ15]' , 'Near infrared spectroscopic determination of diesel fuel parameters using genetic multivariate calibration[Özd08]' and 'Multivariate analysis of near-infrared spectra using the G-programming language[SBB00]'. At last recursive Gaussian process regression and related papers to study could be studied by 'Recursive Gaussian process regression[Hub13]'.

4 Problem Description

The problem statement is the same as that of any topic in predictive modelling i.e. we have a dataset or set of information regarding some of the aspects of the outcome which we want to predict for future cases. The aspects or properties could be numerical or categorical(which has only discrete values). A mathematical model or function is to be approximated using these properties to give correct values of the future outcome. There are many algorithms for this function approximation as listed earlier and are used for different applications. As the degree and complexity of this approximation could be very high and thus very difficult to calculate we use these algorithms to make this task easier.

5 Mathematical Formulation

In probability and statistics, a random variable or stochastic variable is a variable whose possible values are outcomes of a random natural phenomenon.
A probability distribution is a mathematical function that, told in simple terms, could be seen as the distribution of probabilities of occurrence in different possible outcomes in an experiment[Fel08]. A probability distribution is mostly recognized by its mean(μ) and variance(σ).
One such distribution is Gaussian distribution that could be said in simple language as a distribution of probability with maximum density in the mid-region.

One important theorem describing the conditional probability of multivariate Gaussian distribution[HC95] when mean and variance is given as follows:-

Theorem 1(Marginals and Conditionals of an MVN)[HC95]. Suppose $X = (x_1, x_2)$ is jointly Gaussian with parameters

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

Then the marginals are given by

$$p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11})$$

$$p(x_2) = \mathcal{N}(x_2|\mu_2, \Sigma_{22})$$

and the posterior conditional is given by

$$p(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2})$$

where,

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}$$

A Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.[RW06] And also as discussed earlier the main essence of **GPR** is generated due to it being a Bayesian method[Ras06]. In general Gaussian Process is defined as a prior probability distribution over $f(x)$ values such that set of $f(x)$ values at points $x_1, x_2, ..$(given points) jointly have a prior Gaussian distribution[WR06] i.e.

$$f \sim GP(m, K) \quad OR \quad f(X) \sim N(m(X), K(X, X))$$

where,

$$\begin{aligned} m(X) &= E[f(x)] \\ K(x, x^{'}) &= E[(f(x) - m(x))(f(x^{'}) - m(x^{'}))] \\ X &= \left[x_1^T, \ldots, x_n^T\right]^T \text{ where } \{x_1, x_2, ..\} \text{ are given data points} \end{aligned}$$

**Note:- $x^{'}$ in this report is any point in n-dimensional space and not transpose.

The main pivot of this algorithm is its Kernel matrix(could also be seen as covariance matrix in Theorem 1[HC95]). A kernel is basically a function that measures similarity between the points given in the training set or the points plotted. Like for example in this plot the kernel would output much more similarity for the points near to each other and less similarity for the points far away. So in layman's term GPR just estimates $f(x)$ for the unknown points through interpolation from near points. And this is done by choosing appropriate kernel function. Commonly used Kernel Functions are Exponential, Gaussian [RW04]etc.

For our analysis default Kernel used is of squared exponential type:-

$$K(x, x^{'}) = \sigma^2 exp(-\frac{(x - x^{'})^T (x - x^{'})}{2l^2})$$

where,

x and $x^{'}$are any two points in proper range of the variable.
'σ' and 'l' are hyper parameters(generally both taken as 1).

But there are some restrictions on Kernel Matrix[RW06]:-
1) K needs to be symmetric i.e.

$$K(x, x^{'}) = K(x^{'}, x)$$

2) K needs to be positive semi-definite i.e.

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} K(x, x^{'}) g(x) g(x^{'})\, dx\, dx^{'} > 0 \tag{1}$$

for all $g\ \epsilon L_2$(Mercer's Theorem).

6 Experimental Setup

For our analysis we will generate many sample datasets and work on them to check and analyze each and every aspect of this model. Procedure for the same is:

1. Creation of independent variables(X_1 in this case) using random number generator and range setting according to the need.

2. Now assuming the data to be linearly regressed(i.e. y to follow linear relationship with independent variables), we create 'y' variable using the formula $y = \alpha_0 + \alpha_1 x$, where α_0 and α_1 are fixed constants set randomly.

3. For the non-linear data we created 'y' variable using the formula $y = \alpha_0 + \alpha_1 x^2 + \alpha_3 x^3 + \alpha_4 x^6$, where α_i's are fixed constants set randomly. But the formula is changed according to the needs in some situations.

4. We then normalize 'y' variable(as we have assumed earlier) or transform the values to required range. This is done because the model is designed in a similar sort of manner.

7 Solution Methodology

1. Initially we note different points in X axis where the points are given.

2. Then we have to see for the prior similarity between the points which is done using the kernel formula defined earlier.

3. As we have assumed that these points will follow a multivariate Gaussian distribution in itself we note down the mean and variance of the distribution[Bis06].

4. We find each elements of the kernel matrix using the formula for Kernel[RW04] and take μ_*(mean) as 0 for the prior distribution. Also setting the hyper parameter values accordingly.

5. Now we have a set of functions that form a Gaussian distribution and would be able to approximate the future points efficiently.

6. We will use this information(or prior information) to predict for test points. But for this we need a relation to link these new points to the previous points in some manner. This is done using Theorem 1 described earlier which calculates the conditional probability distribution and thus giving the required results.[Fel08]

7. We use the multivariate Gaussian theorem[Fel08] to find the conditional probability distribution of f_{post} using the formula :

$$f_*|f, X, X_* \sim \mathcal{N}(\mu_* + K_*^T K^{-1}(f - \mu), K_{**} - K_*^T K^{-1} K_*)$$

A similar function made in python language could be seen in Appendix.

8. Here we assume μ_*(mean for the f_{post}) also as $[0, 0, 0, \dots]^T$ and Kernel Matrix associating both prior and posterior as K^* having elements in the kernel matrix computed using every possible pair formed by taking one from prior points set and other from posterior points set.

9. Also K^{**} is a kernel matrix which is computed using all pair of points from only posterior points set.

10. Finally using this formula we estimate each new f^* value either by taking only mean as a prediction or we take random samples from this finally formed Normal distribution.

8 Results and Discussion

We have generated different datasets accordingly and compared accuracies of GPR with linear regression under different circumstances. We have normalized the y values before predicting.

The accuracy of Linear model is more in the case when the training and testing data have different ranges as we are effectively changing different ranges to again 0-100 by subtracting lower limit and are using it to predict (Refer Table-1). Similarly, we have generated our results for non-linear dataset where non-linear MSE is calculated using the in-built spline modelling(Refer to Table-2) functions. The major reasoning that could be attached to it is that linearity is most effectively caught by linear models than any other model. But GPR also shows considerable results. For the non-linear case we could infer that non-linearity could also be caught by our model with much ease.

Table 1: Data Ranging for Linear

Range_Training	Range_Test	MSE_Linear	MSE_GPR
0-100	300-400	0.0000207505	0.0007585764
0-100	500-600	0.0000207236	0.0007579788
0-100	(-200)-(-100)	0.0000163694	0.0006755753
0-100	1000-(1100)	0.0000353439	0.0010064371
0-100	(-1100)-(-1000)	0.0000437800	0.0008191817

Table 2: Data Ranging for Non-Linear

Range_Training	Range_Test	MSE_Non-Linear	MSE_GPR
0-100	300-400	0.1585764640	0.3329812111
0-100	500-600	0.5995103201	2.7236277334
0-100	(-200)-(-100)	0.0905403332	0.1803096996
0-100	1000-(1100)	2.7171288100	11.9013627116
0-100	(-1100)-(-1000)	3.2648484358	15.8276890960

For, further analysis we will see the variations in the results by variation with noise that is added to the standard dataset that is generated at the starting. Now, in the case of variation of σ values of the noise given to the y values,we see that as noise increases the error also increases generally but Linear Regression[SL12] shows much better results than GPR in all cases.(Refer Table-3)
For the non-linear case we see that the error shows consistency with the previous results. This could be accounted by saying that as noise varies it gets added to the function values and which in turn adds similar non-linearity to the data.

Table 3: Noise Variation Linear

σ^2	MSE_Linear	MSE_GPR
1	0.0000211595	0.0007288664
10	0.0022484797	0.2237571362
50	0.0083530387	0.2353817968
100	0.0430781442	0.2720123189
400	0.2205892292	0.2268918974
600	0.1906741358	0.2111686014
900	0.2322241757	0.2405578789
1000	0.1084470759	0.1168563747
1200	0.1201096811	0.1219125752
1500	0.1623072601	0.1659302335

Table 4: Noise Variation Non-Linear

σ^2	MSE_Non-Linear	MSE_GPR
1	0.1286545051	0.2210945405
10	0.1370084159	0.2940731075
50	0.2245323190	0.3350846043
100	0.4901541584	0.5510770902
400	0.5572105121	0.6456544222
600	0.4886890714	0.6587072010
900	0.5875701519	0.7630187652
1000	0.5821500071	0.9370150731

In general the raw data we get may contain many outliers. These outliers may deteriorate the performance of the model and thus is essential aspect to consider in model evaluation. We need to see the robustness of the mpdel we are training and then accordingly take steps to deal with it. The outliers in this case are generated by randomly generated x values and corresponding y values in the range of 10 to 11 times that of average x value. From Table-5 we see that when outliers are added in different regions of the data the error for linear regression is approximately 10 times less than in the case of GPR. Thus, we could conclude from this that GPR cannot handle ouliers in an effective manner. In Table-5 outlier sets are different sets of ouliers introduced in the model in different x region. We could also see similar results in the case of

non-linear dataset. For this experiment there is not much difference between the analysis with outliers and without them as there is already non-linearity in the dataset and outlier introduction does not change anything further.

Table 5: Outlier Handling Linear

Outliers Set No.	MSE_Linear	MSE_GPR
1	0.0558568880	0.5664797663
2	0.0491614330	0.5833558234
3	0.0548594275	0.4463948561

Table 6: Outlier Handling Non Linear

Outliers Set No.	MSE_Non_Linear	MSE_GPR
1	0.1985039923	0.6551006174
2	0.4339170792	2.8984659521
3	0.3546802700	2.4239407392

We see from Table-7 that we reach an optimum 'σ' but 'l' has not much effect on error. Also, we see that as we approach near the σ value of the noise, accuracy increases and GPR also may overcome linear regression.[SL12]

Table 7: Parameter Tuning

σ	l	MSE_Linear	MSE_GPR
1	1	0.1611052	0.3894240
10	1	0.1268882	0.3690756
50	1	0.2015352	0.2164591
75	1	0.1718666	0.1671773
100	1	0.2762269	0.3113301
200	1	0.1443215	0.1460400
75	0.4	0.2096775	0.2341971
75	5	0.2582296	0.2704346
75	10	0.1812161	0.1800867
75	50	0.2114793	0.2162537

8.1 Diesel Data Analysis

Data used in this paper was collected by Southwest Research Institute in a project sponsored by the U.S. Army. Eigenvector Research Incorporated provides 784 samples in Near Infrared Spectra of Diesel Fuels.[FWZ15] Some important properties of diesel are explained below.

Cetane Number - It is the measure of ignition characteristics of diesel fuel oil for any compression ignition taking place. It is used by petroleum marketers, engine manufacturers. Also it is used for primary specification measurement relating to matching fuels and engines.[SBB00]

Boiling Point at 50% Recovery - The ASTM method D86-95 covers the purification of diesel fuel and other petroleum based products, done either manually or through automated equipment. A 100-mL sample is filtrated under prescribed conditions. Appropriate and systematic observations of thermometer readings and volumes of condensate are observed, and from these data, the boiling-point-at-50% recovery of the condensate is determined.[SBB00]

Viscosity - The ASTM method D445-94 specifies a method for determining the kinematic viscosity, v, of liquid petroleum products, either it be a transparent or opaque. This is done by measuring the time for a volume of liquid that flows under the influence of gravity through a calibrated glass capillary viscometer. Kinematic viscosity of diesel is reported in units of mm^2s^{-1}. In this report, the kinematic viscosity was written in centiStokes (cSt), where 1 cSt = 10 mm^2s^{-1}. The dynamic viscosity(η) can be calculated by multiplying the measured kinematic viscosity to the density(ρ), of the liquid.[SBB00]

Some of the properties are not measured on some samples, in this case, samples which have missing values were removed.

The spectra of full sample was plot in Fig.1, the x-axis was wavelength, ranged between 750 nm and 1550 nm with a resolution of 2 nm. The z-axis was absorbance.

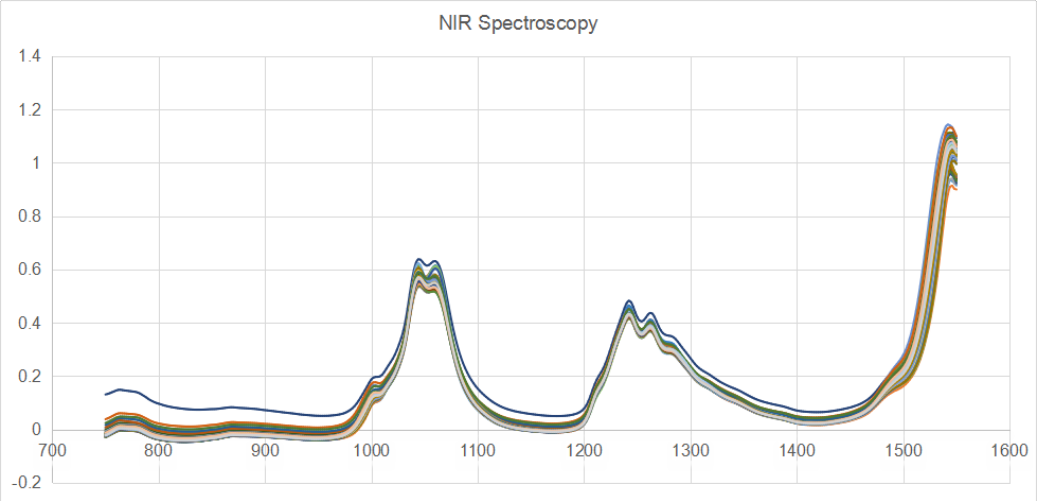

Figure 1: The NIR spectra of full sample set

For our analysis we predicted different aspects of diesel fuel using the NIR wavelengths variations. The results came out to be in par with other experiments done using other algorithms. Main algorithms used apart from GPR are LS-SVM, PLS[FWZ15], genetic multivariate analysis[Özd08], PCR etc. But in most of the papers we have some form of pre processing of data being used to improve the performance of the result. Also, other papers[Özd08][FWZ15][SBB00] referred here uses only 3 types of diesel whereas in this paper we have used all the types making it more difficult to analyze and model. The comparison of corresponding results of GPR and PCA based prediction are as follows:-

Table 8: Diesel data error comparison

Properties	$(\text{GPR}_{rmse}/Avg.Value)$	$(\text{PLS}_{rmse}/Avg.Value)[FWZ15]$
BP50	0.1393631048	0.14622
CN	0.5510231537	0.14858
Viscosity	0.1332196107	0.0745

We observe here that the accuracy measure for BP50 for our method gives better results than the PCA based pre processed analysis[FWZ15]. Also, we see from Figure-2, for other variables we have comparable outputs and the algorithm is able to catch most of the non-linearity with much ease. We

could therefore use it for these type of datasets with so much variation.

9 Conclusion and Future Work

The algorithm discussed in this paper is universal in nature and thus could be used in any applications. The results also prove the effectiveness of this algorithm. But there could be more improvement in the results through more data pre-processing as seen in other related papers. Apart from that, the outcomes are more than satisfactory to be used in real life applications. One such detailed discussion has been done on the NIR spectroscopy analysis of diesel fuels. The good agreement of the method used in this paper demonstrates that the method could be done online without any pre-treatment of complex samples compared with chemical methods. Besides, properties could be analyzed simultaneously, and the whole analysis process only takes seconds. One important conclusion could be seen in the results is that there is significant difference in RMSE of GPR when compared to other papers. This is certainly due to the special type of pre-processing involved while improving the results. Future scope of this research is discussed below.

9.1 Tuning Hyperparameters

We observe here that while using kernel function we assumed that both 'σ' and 'l' used as hyperparameters were taken as 1. But this was only used for simplicity. If we try to optimize these hyperparameters, we see that while differentiating the log likelihood equation w.r.t these hyperparameters we get K(kernel) also in it and hence we need to calculate K^{-1} and $dK/d\theta$ for each iteration in gradient based optimizer which is not feasible(making complexity very high)[RW06].

Thus many techniques have been used to overcome this pitfall such as maximizing log likelihood by conjugate gradients method or by using various sampling techniques like HMC(hybrid Monte Carlo sampling), MCMC(Markov chain Monte Carlo)[DFN02], Bayesian optimization based search[SLA12], grid Search. But all these methods are also very slow in performance and usually find local optimum values.

By observation we see that the when values of 'σ' and 'l' are near to the

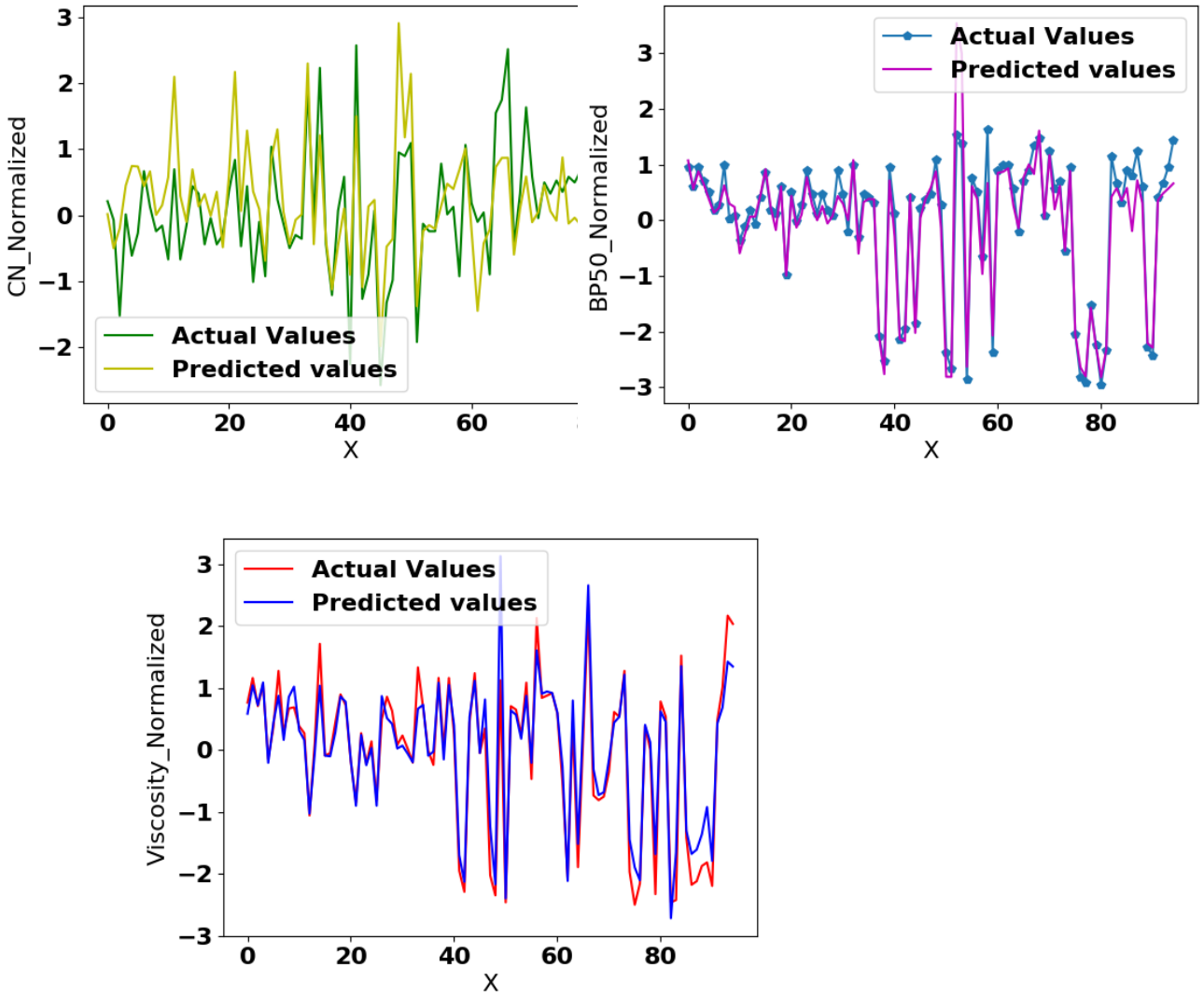

Figure 2: Diesel Data Results

'σ' value of the the noise it gives the best predictions. So, for future the method could be to first estimate that noise and then sample around that using any of the above mentioned sampling techniques to reach optimum result.

9.2 Recursive Gaussian Process Regression[Hub13]

We face many problems while performing the experiment of diesel properties observation. Such as it takes long hours for getting a set of observation values, and for getting 784 set of observations we need 784 days. Also, as we go on doing experiments the setup and accuracy of measurement of instruments changes, thereby affecting our observations. This impact could be seen in the difference in results of Cetane Number prediction. This is where this algorithm could help. We could use recursive Gaussian process regression[Hub13] and use previous day results to again train(or change the parameters) for the current day making it more appropriate. Thus, in future we hope to incorporate an online algorithm inspired by this to deal with this problem.

References

[HC95] Robert V Hogg and Allen T Craig. *Introduction to mathematical statistics.(5"" edition)*. Upper Saddle River, New Jersey: Prentice Hall, 1995.

[SBB00] Olusola O Soyemi, Marianna A Busch, and Kenneth W Busch. "Multivariate analysis of near-infrared spectra using the G-programming language". In: *Journal of chemical information and computer sciences* 40.5 (2000), pp. 1093–1100.

[DFN02] Petros Dellaportas, Jonathan J Forster, and Ioannis Ntzoufras. "On Bayesian model and variable selection using MCMC". In: *Statistics and Computing* 12.1 (2002), pp. 27–36.

[RW04] Carl Edward Rasmussen and Christopher KI Williams. "Gaussian processes in machine learning". In: *Lecture notes in computer science* 3176 (2004), pp. 63–71.

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Ras06] Carl Edward Rasmussen. "Advances in Gaussian processes". In: *Advances in Neural Information Processing Systems* 19 (2006).

[RW06] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. Vol. 1. MIT press Cambridge, 2006.

[WR06] Christopher KI Williams and Carl Edward Rasmussen. "Gaussian processes for machine learning". In: *the MIT Press* 2.3 (2006), p. 4.

[Fel08] Willliam Feller. *An introduction to probability theory and its applications*. Vol. 2. John Wiley & Sons, 2008.

[Özd08] Durmuş Özdemir. "Near infrared spectroscopic determination of diesel fuel parameters using genetic multivariate calibration". In: *Petroleum Science and Technology* 26.1 (2008), pp. 101–113.

[SL12] George AF Seber and Alan J Lee. *Linear regression analysis*. Vol. 936. John Wiley & Sons, 2012.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems*. 2012, pp. 2951–2959.

[Hub13] Marco F Huber. "Recursive Gaussian process regression". In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 3362–3366.

[FWZ15] Fei Feng, Qiongshui Wu, and Libo Zeng. "Rapid analysis of diesel fuel properties by near infrared reflectance spectra". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 149 (2015), pp. 271–278.