Processing synthetic seabed logging (SBL) data using Gaussian Process regression

Muhammad Naeim Mohd Aris, Hanita Daud and Sarat Chandra Dass
Fundamental and Applied Sciences Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia

Email: muhammad_naeim@yahoo.com

Abstract. This paper presents a study on processing one dimensional (1D) synthetic seabed logging (SBL) data which were generated through Computer Simulation Technology (CST) software using Gaussian Process Regression (GPR). Seabed Logging (SBL) is an application of electromagnetic (EM) wave emitted from a controlled source to discover hydrocarbon-saturated layers beneath the seabed. In this paper, GPR is proposed as the processing tool to provide any additional information for SBL application. GPR is able to provide predicted mean values and uncertainty measurement in terms of ± standard deviation. The procedures of regressing Gaussian Process (GP) are described thoroughly in this paper. Squared exponential (SE) is chosen as the covariance function used in the GPR. SE covariance function is capable of producing smooth and infinitely differentiable of predicted functional estimates. Log-marginal likelihood is then optimized in order to infer the hyper-parameters involved in the SE covariance function. For model validation, mean square error (MSE) is calculated and observed to determine the reliability of the GPR model in the synthetic SBL data. This shall give an indication that GPR is an appropriate tool for processing nonlinear SBL data with uncertainty quantification and low MSE.

1. Introduction
Gaussian Process (GP) is a statistical model that can be exercised for Bayesian supervised learning. According to [1], there are two tasks involved in supervised learning which are regression and classification. However, this paper only focuses on Gaussian Process regression (GPR) since the estimation is for continuous dependent variable (magnitude of E-field). GPR puts directly a prior on the function space, \( f \), without parameterizing the function. Due to that, GP can be fitted in many different forms of data without any parameter alteration as long as covariance function is defined to each type of data. Processing 1D synthetic SBL data using GPR model could provide beneficial information for computational and visualization of EM wave application.

SBL application is widely used in deep offshore hydrocarbon (HC) exploration since last decades. This frequency domain technique exploits EM wave to make reliable measurements of subsurface electrical resistivity underneath the seabed. Marine controlled-source electromagnetic (CSEM) technique has been considered as a necessary tool to interpret the acquired seismic imaging data in order to reduce the...
exploratory risk and ambiguities [2]. This is because seismic imaging technique fails to distinguish the fluid content in the potential reservoir whether it is hydrocarbon- or water-filled reservoir [3]. SBL survey generates huge amount of noisy data [4]. Thus, flexible model that is able to provide the uncertainty quantification is required in the data processing in order to handle the noise.

2. Literature Reviews
This section provides brief understanding on GPR, SBL application and Computer Simulation Technology (CST) software.

2.1. Gaussian Process Regression (GPR)
Basically, Gaussian Process regression (GPR) model is a distribution over function. This probabilistic and non-parametric model is very flexible [5]. GPR model has been known as a machine of probabilistic kernel and has been applied in many scientific and engineering application areas which comprises machine learning, geo-statistics, electronics etc. [6]. The main advantage of using GPR is, it can provide variance for predicted mean distribution as a description of uncertainty measurements. Over-fitting problem also can be avoided in the predictions since it follows Bayesian interpretation setting [6].

Theoretically, GPR is defined as a collection of random variables such that any of the variables is distributed by Gaussian distribution [7]. A GP is fully specified by mean function, \( m(x) \), and the covariance function, \( k(x,x') \). Each random function, \( f(x) \) is Gaussian distributed, \( f(x) \sim N(m(x),k(x,x')) \) where \( x \in D \). The \( n \) dataset, \( D = \{ (x_i,f_i) \mid i = 1,2,\ldots,n \} \) also is said to follow GP distribution. Detail explanations on GPR will be described in-depth in methodology section.

2.2. Seabed Logging (SBL) Application
Seabed Logging (SBL) application is usually employed in deep offshore environment for locating hydrocarbon reservoirs beneath the seabed using EM wave. This application assumes that hydrocarbon has higher electrical resistivity compared to its surrounding hosts [8]. Researcher in [9] stated that hydrocarbon has electrical resistivity of 30-500\( \Omega \text{m} \) compared to seawater and sedimentary rocks which are 0.5-2\( \Omega \text{m} \) and 1-2\( \Omega \text{m} \), respectively. SBL application has been successfully applied in various geological physical structures of oceans such as offshore of West Africa, Norway, Brazil and Gulf of Mexico [10]. Figure 1 shows the schematic diagram of SBL application during survey process.

![Figure 1. Representation of SBL survey](image-url)
During SBL survey, EM wave is emitted from a mobile source or known as Horizontal Electrical Dipole (HED) transmitter and EM receivers record the returned EM signals [11]. HED transmitter is usually elevated at 30-40m above the seabed. According to [12], in order to generate higher wavelength that plays role as the distance function, low frequency EM wave (0.01-10Hz) is exercised in the survey. Conceptually, electric field (E-field) of EM energy is slowly attenuated in high resistive medium due to skin depth effect compared to high conductive medium (low resistivity).

2.3. Computer Simulation Technology (CST) Software
Well propagation of EM signal can provide good information about the structure of the subsurface beneath the seabed. [13] mentioned that Maxwell’s equation is required to predict the propagation of low frequency EM wave. CST software is able to discretize the Maxwell’s equation in order to probe the resistivity contrast [14]. CST EM Studio uses finite-integration technique (FIT) to solve the equations in finite calculation domain in grid cell [15]. It can be in hexahedral or tetrahedral mesh elements. Researchers in [16] used mesh element of tetrahedral in CSEM modeling.

3. Methodology
This section explains the methodologies involved in this study which includes the procedures of the GPR.

3.1. SBL Modeling and Synthetic Data Acquisition
Synthetic SBL data are generated through Computer Simulation Technology (CST) software. The simulation model is designed in rectangular prism where it consists of air, seawater, sediment (overburden and under burden layers) and hydrocarbon. The overburden thickness of the SBL model is fixed at 500m. 270m long transmitter is positioned at the center of the model at 30m above the seabed. Frequency of EM wave used in this study is 0.125Hz. The volume of the SBL model is $10 \times 10 \times 5 \text{ km}^3$ and for the hydrocarbon is $10 \times 5 \times 0.2 \text{ km}^3$. Every layers involved in the simulation model have unique permittivity, electrical conductivity and thermal conductivity. Specific physical properties for each layer are followed as in [4]. Figure 2 shows the simulation model of SBL application.

3.2. Gaussian Process Regression (GPR)
Researchers in [5] mentioned that basic GPR used a zero mean function as generalization purposes. Zero mean prior does not imply zero mean in posterior distribution. Hence, GP distribution on function, $f$, can be written as $f(x) \sim N(0, k(x, x'))$. This paper assumes to have a relationship, $Y = f(X)$, where training set input (offset: distance of EM signals travel from the source to the receiver),
\( X = \{ x_i \mid i = 1,2,\ldots,n \} \), and the corresponding set output (magnitude of E-field), \( Y = \{ y_i \mid i = 1,2,\ldots,n \} \). Gaussian distribution over the corresponding output with zero mean function can be defined as 
\[
[y(x_1), y(x_2), \ldots, y(x_n)]^T \sim N(0, K).
\]

\[
K = \begin{bmatrix}
k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\
k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\
\vdots & \vdots & \ddots & \vdots \\
k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n)
\end{bmatrix}
\]

\( K \) is the \( n \times n \) covariance matrix with elements of covariance function in training inputs.

The smoothness of GPR model is influenced by the correlation behavior in the covariance function \([7]\), and the reliability of GPR model relies on the performances of the covariance function. Therefore, squared exponential (SE) is selected as the covariance function as it is a popular choice among other covariance functions in kernel machine fields \([1]\). The SE covariance function can be defined as

\[
k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2l} \left( \frac{x-x'}{l} \right)^2 \right)
\]

where \( \sigma_f \) and \( l \) represent signal variance and characteristic of length scale, respectively. Hyperparameters, \( \theta = \{\sigma_f, l\} \), must always be positive and must be sensibly chosen in order to get the best correlation between data points. This corresponds to the optimization of the log-marginal likelihood function. The negative log-marginal likelihood \([17]\) is defined as

\[
\log p(Y \mid X, \theta) = -\frac{1}{2} \left( \log \det K + n \log(2\pi) \right)
\]

where \( p(Y \mid X, \theta) \) is distributed by \( N(0, K) \).

GPR is able to predict expected output at any desired input or also called as test set input, \( X^* = \{ x_i^* \mid i = 1,2,\ldots,m \} \) and the prior joint distribution for the random variables collection can be written as
\[
[y(x_1), y(x_2), \ldots, y(x_n), Y^*] \sim N(0, \kappa) \text{ where } \kappa = \begin{bmatrix} K & K^T \\ K^* & K_{**} \end{bmatrix}.
\]

\( K_{**} \) and \( K_{*} \) represent the training-test set of covariance and test set covariance, respectively.

\[
K_{**} = [k(x^*, x_1) \ k(x^*, x_2) \ \cdots \ k(x^*, x_n)]
\]

\[
K_{*} = k(x^*, x^*)
\]

By following multivariate Gaussian theorem, the conditional distribution or posterior distribution \([1]\) can be derived as
This posterior distribution is exploited in order to predict the corresponding test output [18] which are the estimated mean, \( m_* \), and the estimated variance, \( k_* \). Both equations are essential for GPR prediction.

\[
\begin{align*}
Y^* | Y & \sim N(K_*K^{-1}Y, K_* - K_*^{-1}K_*^T) \\
\end{align*}
\]

\( Y_* = K_*K^{-1}Y \)

\( k_* = K_* - K_*^{-1}K_*^T \)

3.3. Calculating Mean Square Error (MSE)

Mean square error (MSE) is one of many prediction error tools to quantify the differences between the true values and the predicted values. The differences are observed to evaluate the performance of the prediction. According to [9], the measured error is the amount by which the estimator, \( y_i \), differs from the quantity that is estimated, \( y_i^* \). MSE of \( y_i \) with respect to \( y_i^* \) is defined as follow where \( n \) is the total number of data,

\[
MSE = \frac{1}{n} \sum (y_i - y_i^*)^2
\]

4. Results and Discussions

GPR was fitted by using Gaussian Process Machine Learning (GPML) toolbox in MATLAB code written by [19]. First five data, which consist four training set and one test input, were selected for the purpose of the GP regression. The steps to estimate the target test output (log10 of magnitude of E-field), \( y_* \) at the test input (offset), \( x_* = 239926 \), are elaborated thoroughly as below.

Four observations, \( y(X_i) \) at \( X = [2397.26,2398.26,2400.26,2401.26] \) were selected as the training set input for this demonstration where the corresponding outputs were normally distributed, \( [-5.4524, -5.4530, -5.4541, -5.4547]^T \sim N(0, K) \). \( y \) and \( x \) represented the log10 of the magnitude of E-field and offset, respectively.

At first, the hyper-parameters were chosen randomly from the prior distribution. Equation (3) was numerically minimized using the initial hyper-parameters, and an estimate of hyper-parameters was obtained from the process. This process is called as optimization. The optimization was repeated and the estimate with the smallest negative log-marginal likelihood was selected as the optimal estimate of the hyper-parameters. The optimal hyper-parameters used in this demonstration; \( \sigma_v = 6.63671909 \) and \( l = 1230901914 \). Then, matrix (1) was constructed using equation (2).

\[
K = \begin{bmatrix}
44.04604298 & 44.04604283 & 44.04604167 & 44.04604065 \\
44.04604283 & 44.04604298 & 44.04604240 & 44.04604167 \\
44.04604167 & 44.04604240 & 44.04604298 & 44.04604283 \\
44.04604065 & 44.04604167 & 44.04604283 & 44.04604298
\end{bmatrix}
\]
Next, the prior joint distribution for the training-test observations can be written as 
\[
\begin{bmatrix} -5.4524, -5.4530, -5.4541, -5.4547 \end{bmatrix} \sim N(0, \kappa). 
\]
Equation (2) also was used to calculate matrices (4) and (5) which are the training-test covariance and the test set covariance. The covariance matrices; 
\[
K_e = \begin{bmatrix} 44.04604240 & 44.04604283 & 44.04604283 & 44.04604240 \end{bmatrix} \quad \text{and} \quad K_* = [44.0460429]. 
\]
Finally, from equations (7) and (8), the estimated mean (log10 of magnitude of E-field at \( x^* = 2399.26 \)), \( m_e = -5.45354935 \) and the estimated variance, \( k_e = 2.4158E-13 \) were calculated.

In order to determine the reliability of the GPR model, the estimated target test output was then validated by calculating the MSE, as equation (9), between the true output value at \( x^* = 2399.26 \) and the predicted value. The calculated MSE is \( 1.4086E-13 \). This implies that the model has a very small error in the prediction and the calculated MSE is highly accepted since it is approaching to zero. However, it should be noted that the MSE was calculated for only one data point. Thus, MSE for higher number of data is discussed later. Figures below show the plots for fitted GPR model on the dataset. Since the error bar is hardly to be seen, this paper provides the figure in zoom-in scale to clearly see the 95% confidence interval at the test data.

![Figure 3. Fitted 1D GPR model](image-url)

![Figure 4. 95% CI at \( x^* = 2399.26 \)](image-url)

In figures above, the x-axis represents the offset, and the y-axis represents the magnitude of E-field in log10 scale. The purpose of using log10 scale is to make the data interpretable. Based on Figure 3, the round symbols denote the training data and the zoomed-in cross symbol denotes the test observation, \( (x^* = 2399.26, y^* = -5.453549355) \). The vertical line in Figure 4 designates \( \pm 2\sigma \), which represents the uncertainty quantification for the test data. Based on Figure 4, the actual point is still within the vertical line (95% confidence interval) of the predicted mean. This implies that the prediction is fit well in the dataset.

This paper also provides the GPR for 1199 data to determine whether GPR can fit in huge amount of data or not. The dataset consisted 766 training set and 433 test set data based on suggestion by [20] where about one-third of the dataset should be considered as the test data. The hyper-parameters were inferred by multivariate optimization algorithm as elaborated before. Figure 5 shows the GPR model which was fitted to the huge number of data.
The grey-colored region designates ±2σ (95% confidence interval) for the predicted test data. The MSE between the 433 actual and predicted test data was calculated, and the calculated MSE is 2.7044E-05. This information gives a good implication that GPR model also can fit well in huge amount of SBL data with 95% confidence interval.

5. Conclusions
This paper reveals that Gaussian Process regression (GPR) can fit well in 1D synthetic SBL data with low MSE. The outputs are also predictable with uncertainty quantification which is 95% confidence interval. Processing the 1D SBL data by using GPR must be taken into consideration as future work. GPR has a good potential as data processing tool to provide beneficial hydrocarbon information especially to de-risk the hydrocarbon exploration for SBL application.

Acknowledgement
The authors would like to express their appreciation and gratitude to those who have assisted towards completing this paper work especially to Universiti Teknologi PETRONAS funding under Grant YUTP (Cost Centre: 0153AA-E04).

References
[1] Rasmussen C E and Williams C K I 2006 Gaussian Processes for Machine Learning vol 14 Cambridge The MIT Press pp 1-128
[2] Kumar A and Lie J E 2008 Sea bed logging – direct hydrocarbon detection technique in offshore exploration 7th International Conference & Exposition on Petroleum Geophysics
[3] Feather K 2007 The rapid adoption of seabed logging Scandinavian Oil and Gas Magazine 5/6 pp 37–38
[4] Mukhtar S M, Daud H and Dass S C 2015 Prediction of hydrocarbon using Gaussian process for seabed logging application Procedia Computer Science 72 pp 225–232
[5] Liu D, Pang J, Zhou J, Peng Y and Pecht M 2013 Prognostics for state of health estimation of lithium-ion batteries based on combination Gaussian process functional regression
Microelectron. Reliab. 53 pp 832–839

[6] Harari O and Steinberg D M 2014 Optimal designs for Gaussian process models I via spectral decomposition J. Stat. Plan. Inference 154 pp 87–101

[7] Srinivas N, Krause A, Kakade S M and Seeger M W 2012 Information-theoretic regret bounds for Gaussian process optimization in the bandit setting IEEE Trans. Inf. Theory 58(5) pp 3250–3265

[8] Wiik T, Hokstad K, Ursin B and Møttchersd L 2013 Joint contrast source inversion of marine magnetotelluric and controlled-source electromagnetic data Geophysics 78 pp E315–E327

[9] Daud H, Yahya N, Sagayan V and Talib M 2011 A scaled experiment 2 for verification of SPLINE Interpolation Technique for sea bed logging method in Proceedings - 2011 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2011 pp 80–85

[10] Chiadikobi K C, Chiaghanam O I, Omoboroiwo A O, Etukudoh M V and Okafor N A 2012 Detection of hydrocarbon reservoirs using the controlled-source electromagnetic (CSEM) method in the ‘beta’ field, deep water offshore Niger Delta, Nigeria Int. J. Sci. Emerg. Technol. 3 pp 7–18

[11] Li Y and Key K 2007 2D marine controlled-source electromagnetic modeling: Part 1 — An adaptive finite-element algorithm Geophysics 72(2) p WA51

[12] Daud H, Yahya N, Sagayan V and Talib A M 2010 Magnitude vs low frequency EM waves for sea bed logging applications 2010 IEEE Asia-Pacific Conference on Applied Electromagnetics, APACE 2010 – Proceedings

[13] Rauf M, Yahya N, Nyamasviswa T E, Ansari A, Shafie A and Nahar N 2015 Prediction of double stacking hydrocarbon using marine controlled source electromagnetic method 2014 IEEE Asia-Pacific Conf. Appl. Electromagn. APACE 2014 - Proceedings pp 99–102

[14] Yahya N, Nasir N, Akhtar M N, Kashif M and Mohd Zaid H 2012 Modeling of antenna for deep target hydrocarbon exploration J. Electromagn. Anal. Appl. 4 pp 30–41

[15] Canova A, Dughiero F, Fasolo F, Forzan M, Freschi F, Giaccone L and Repetto M 2009 Simplified approach for 3-D nonlinear induction heating problems IEEE Trans. Magn. 45(3) pp 1855–1858

[16] Cai H, Xiong B, Han M and Zhdanov M 2014 3D controlled-source electromagnetic modeling in anisotropic medium using edge-based finite element method Comput. Geosci. 73 pp 164–176

[17] Gibson N P, Aigrain S, Roberts S, Evans T M, Osborne M and Pont F 2011 A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy Mon. Not. R. Astron. Soc. 419 pp 2683–2694

[18] Saha S, Saha B, Saxena A and Goebel K 2010 Distributed prognostic health management with Gaussian process regression Aerospace Conference, 2010 IEEE pp 1–8

[19] Rasmussen C E and Nickisch H 2017 The GPML Toolbox version 4.1 pp 1–49

[20] Pal M and Deswal S 2010 Modelling pile capacity using Gaussian process regression Comput. Geotech. 37 pp 942–947