Bayesian Linear Regression Model for Curve Fitting

Michael Li

Bayesian linear regression model for curve fitting

Michael Li

CIS and School of Engineering and Technology,
Central Queensland University
Rockhampton, Queensland 4701, Australia

m.li@cqu.edu.au

Abstract. This article describes a Bayesian-based method for solving curve fitting problems. We extend the basic linear regression model by adding an extra linear term and incorporating the Bayesian learning. The additional linear term offsets the localized behavior induced by basis functions, while the Bayesian approach effectively reduces overfitting. Difficult benchmark dataset from NIST and high-energy physics experiments have been tested with satisfactory results. It is intriguing to notice that curve fitting, a type of traditional numerical analysis problem, can be treated as an adaptive computational problem under the Bayesian probabilistic framework.

Keywords: Bayesian learning, RBF, curve fitting, stopping power

1 Introduction

The goal of curve fitting is to find a simple analytical function that best fits a set of data points. The best fit means that a certain error measure (such as mean squared error) should be minimized. Curve fitting is a challenging task because a single parametrized function may often not be able to represent a complicated curve due to the complexity of scattered data distribution. In addition, overfitting is a common problem where over-matching numerically the requirement for the fit causes a severe deviation of the data trend. In terms of the terminology of neural computing, the model for fitting is over-trained and leads to a poor generalization performance. To improve the fitting result, a linear combination of a set of basis function can be considered to replace the single parameterized function and efficient methods to prevent overfitting should be introduced. The typical efficient approaches to prevent overfitting includes using regularization and Bayesian prior. The former adds a penalty term in the objective function while the latter is to apply Bayesian probabilistic model to reduce overfitting.

Bayesian approach resolves the overfitting problems in curve fitting or regression analysis with two primary elements: (i) A full probabilistic description of the computational model; and (ii) Use of Bayes' theorem. The former consistently deals with uncertainties for data model and its parameters in terms of probability distribution, while the latter is used to make an information inference related to a learning process

from data to model through a conditional probability relationship. Denoting the observed data by D and the model parameter vector by $\mathbf{w}$ with assuming a prior probability distribution p($\mathbf{w}$), Bayes' theorem can be expressed into the following form

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})\, p(\mathbf{w})}{p(D)}$$

where p(D|$\mathbf{w}$) is called the likelihood function that evaluates the probability of the observed data for a given parameter $\mathbf{w}$. The term p(D) can be obtained by integrating p(D|$\mathbf{w}$)p($\mathbf{w}$) over all $\mathbf{w}$ and it can be viewed as a normalization constant. The term p($\mathbf{w}$|D) is known as the posterior probability distribution. Essentially Bayes' theorem represents a learning process in which it transforms the prior knowledge to the posterior distribution with knowledge updates and highlights the fact that we have learned about the validity of the model parameter from consideration of the observed data.

Curve fitting based on Bayesian inference and linear regression models has been studied by several authors [1-4]. Most of the existing research in Bayesian curve fitting highlight using piecewise polynomials. Motivated by Bayesian reasoning, Denison et al. [1] proposed a method by using a series of piecewise polynomial for fitting a variety of curves. In any method with piecewise functions, the knot selection always is a crucial issue. Instead of directly selecting them, the number and the locations of the piecewise polynomials were modelled as parameters to be inferred in their method; a joint probability distribution was first built over them. Then the reversible jump Markov Chain Monte Carlo (MCMC) technique [3] was used to compute the posteriors. The presented method has achieved a good approximation for some continuous and smooth functions, even for those rapid varying curves. However the simulation results indicated that it was a computation expensive method, with the running time of a single fitting task up to 30 minutes on a SUN SPARC 5 workstation [1]. Dimatteo et al. [2] extended Denison's method by developing a regression model called Bayesian adaptive regression splines. In their model, the free cubic splines were used as a set of basis functions and the number of knots and their positions were allowed to be free parameters that were determined from the data. Dimatteo's model constructed a marginalised chain on the knot number and locations with providing methods for inference on the regression coefficients. Poisson priors on the number of knots were adopted. Their approach also applied the reversible jump Metropolis-Hasting MCMC simulation on the parameter pair - the number of knot and their locations. Additionally, Dimatteo's model incorporated an important locality heuristic observation made by Zhou & Shen [5], which efficiently aided to place knots close to existing knots in order to deal with rapid changes. It has been reported that Dimatteo's method presented more accurate estimates for a certain group of test functions. The main limitation may be that their test functions were mainly from an exponential family distribution. More recently curve fitting based on Bayesian quantile regression has received growing attention [4,6-7]. As a robust statistical model, Bayesian quantile regression provides an efficient alternative to the ordinary mean regression, particularly when the measured data contain a large amount of outliers. Chen and Yu [4] described a general approach for nonparametric quantile curve fitting incorporated with Bayesian inference. As usual

Chen and Yu's method performed quantile regression curve fitting using piecewise polynomials with the unknown number of knots and their locations as input parameters to be inferred. They adopted the asymmetric Laplace distribution as the likelihood function, which exhibited more flexibility. As this type of likelihood function introduces an extra scale parameter, it speeds up the convergence of the implementation of MCMC algorithm. This type of likelihood function also allows one to approximate the marginal likelihood ratio of the number of knots and their locations, which is an important factor in the inference for deciding the accept/reject probability. Although Chen and Yu's method was competitive in accuracy and robustness in performing challenging fitting tasks, their approach didn't produce a simple and universal empirical fitting formula for further applications.

In this article, we propose a new method that incorporates Bayesian probabilistic inference in the RBF regression model for curve fitting. In particular, an additional linear term in RBF model has been introduced for a better approximation. Our approach will be utilized to investigate a few benchmark examples and subsequently is applied to fit experimental data from high energy physics measurements where stopping power curves of oxygen projectiles in elemental target materials carbon, silicon and gold are studied. The accuracy of stopping power data has significant influence in two application areas - ion beam analysis technique and radiation therapy [8].

The organization of this paper is as follows. In Section 2, the proposed method is described in details. Next, the benchmark numerical examples are tested, and computer simulation results for stopping power data are discussed in Section 3. Finally, Section 4 concludes the paper.

2 Bayesian probabilistic regression model for curve fitting

In Bayesian data analysis, a key concept is uncertainty. Statistically each value of the observed quantities inevitably falls in a small uncertain range. This mainly arises from measurement errors or noises. Similarly, values of parameters of a statistical computational model may also be in uncertainty, due to the finite size of data set to derive them. The best way for dealing with uncertainties is to use probabilistic modelling, in which both data and model parameters are analytically treated as random variables and their uncertainties are quantified by a probability distribution. Consider a general regression problem where the input variable is a vector $\mathbf{x}$, the target variable is a scalar denoted by t, and an N-points sample data set $\{\mathbf{x}_i, t_i\}_{i=1}^N$ is given. The regression problem that fits the data set $\{\mathbf{x}_i, t_i\}_{i=1}^N$ to an underlying function can be defined as follows:

$$t_i = y(\mathbf{x}_i; \mathbf{w}) + \varepsilon \qquad \text{i=1,…,N} \tag{1}$$

where ε denotes the random error, and $\mathbf{w}$ denotes a vector of all adjustable parameters in the model. Under the linear regression model, the model function $y(\mathbf{x};\mathbf{w})$ is a linearly-weighted sum of M fixed basis functions $\varphi_j(\mathbf{x})$,

$$y(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) \tag{2}$$

The random error in data can be assumed to be a zero-mean Gaussian noise with variance β^{-1}:

$$\varepsilon \sim \mathcal{N}(0, \beta^{-1}) \tag{3}$$

Statistically it is a reasonable hypothesis that noises in data are Gaussian, since the underlying mechanisms generating physical data often include many stochastic processes while the central limit theorem of the probability theory reveals that the summation of many random processes tends to have the normal distribution. Hence from equations (1) and (3), the target variable t is a random variable and its conditional probability upon $\mathbf{x}$ and $\mathbf{w}$ satisfies a normal distribution with a mean equal to y($\mathbf{x}$;$\mathbf{w}$) and variance β^{-1}. It can be expressed as

$$p(t \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \tag{4}$$

As each data point is drawn independently and identically and its probability obeys the distribution of equation (4), the likelihood function of the entire dataset $\{\mathbf{x}_i, t_i\}_{i=1}^{N}$ is the product of the probability of each point occurrence and it is given by

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{i=1}^{N} \mathcal{N}(t_i \mid y(\mathbf{x}_i, \mathbf{w}), \beta) \tag{5}$$

It should point out that from equation (5) and later, we refer to $\mathbf{x}=[\mathbf{x}_1 \ldots \mathbf{x}_N]^T$ and $t=[t_1 \ldots t_N]^T$ as the entire data set $\{\mathbf{x},\mathbf{t}\}$ for convenient notations. It is possible to make a point estimate on the model parameter $\mathbf{w}$ by using the maximum likelihood (ML) estimate in equation (5). However the ML method often leads to overfitting data. To control the model complexity, a prior distribution over $\mathbf{w}$ is introduced. For simplicity, we add an isotopic Gaussian distribution of the form

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I}) \tag{6}$$

where $\mathbf{I}$ is the unit matrix, and α is termed as the hyperparameter of model.

The purpose of curve fitting is to predict the corresponding value t^* of the target variable for a new test point $\mathbf{x}^*$, given the existing sample set $\{\mathbf{x},\mathbf{t}\}$. Therefore it is necessary to evaluate the probability distribution of the predictive t^* i.e. p(t^*|$\mathbf{x}^*$,$\mathbf{x}$,$\mathbf{t}$). In a fully Bayesian treatment of the probabilistic regression model, in order to make a rigorous prediction for a new data point, it requires us to integrate the posterior probability distribution with respect to both the model parameter and hyperparameters. This is because the complete marginalization procedure would make effectively averaging over all different possible solutions corresponding to the individuals ($\mathbf{w}$, α, β). However, the triple integration for a complete marginalization is analytically intractable. As an approximate scheme, the practical Bayesian treatment assumes that the hyperparametrs α and β are known in advance. With this assumption, the expression of predictive distribution p(t^*|$\mathbf{x}^*$,$\mathbf{x}$,$\mathbf{t}$) can be derived through marginalizing over $\mathbf{w}$ [9],

$$p(t^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{t}) = \int p(t^* \mid \mathbf{x}^*, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)\, d\mathbf{w} \tag{7}$$

By using the Bayesian theorem, the posterior distribution p(**w**|**x**,**t**,α,β) can be written as

$$p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w} \mid \alpha) \tag{8}$$

Substituting (4), (5), (6), and (8) into (7), after a series of algebraic manipulations, the predictive distribution p(t^*|**x***,**x**, **t**) can be simplified as a normal distribution

$$p(t^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t^* \mid m(\mathbf{x}^*), s^2(\mathbf{x}^*)) \tag{9}$$

where m and s^2 are the mean and variance of the predictive distribution of t^*, they are given by

$$m(\mathbf{x}^*) = \beta \phi(\mathbf{x}^*)^T \mathbf{S} \sum_{i=1}^{N} \phi(\mathbf{x}_i) t_i \tag{10}$$

$$s^2(\mathbf{x}^*) = \beta^{-1} + \phi(\mathbf{x}^*)^T \mathbf{S} \phi(\mathbf{x}^*) \tag{11}$$

Here the matrix **S** is given by

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{i=1}^{N} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T \tag{12}$$

From the above inference, the posterior probability distribution of the predictive value of the target variable has been derived. In the statistical sense, mean characterizes the central location in a set of points, where the highest probability event occurs for a normal distribution. Hence the mean value $m(\mathbf{x}^*)$ obtained from the predictive distribution equations (10) and (12) is the best approximation to the predictive t^*. It represents the predictive value of the target variable at the new test point **x***.

Under the linear regression model, the regression function y(**x**;**w**) is linear to model parameter **w**, while it is nonlinear to input variable **x**. As you have seen in equation (2), it can be expanded by a set of basis functions. In our previous study [10], we proposed to add a linear term to make a global correction for Gaussian radial basis function which suffers from the localized effect. With adding the extra linear term, the model regression function becomes the following form,

$$y(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{M} w_i \varphi(\| \mathbf{x} - \boldsymbol{c}_i \|) + a\mathbf{x} + b \tag{13}$$

where c_i is the center parameter governing the location of the basis function in the input space, a is the linear coefficient, and b is the constant term. By using the matrix notation, the expressions of (10)-(12) can be re-written as concise matrix forms,

$$\mathbf{m} = \beta \mathbf{S} \boldsymbol{\Phi}^{\mathbf{T}} \mathbf{t} \tag{14}$$

$$s^2 = \beta^{-1} + \phi(\mathbf{x}^*)^{\mathbf{T}} \mathbf{S} \phi(\mathbf{x}^*) \tag{15}$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi} \tag{16}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_1(r_1) & \varphi_2(r_1) & .... & \varphi_N(r_1) & x_1 & 1 \\ \varphi_1(r_2) & \varphi_2(r_2) & .... & \varphi_N(r_2) & x_2 & 1 \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ \varphi_1(r_M) & \varphi_2(r_M) & .... & \varphi_N(r_M) & x_M & 1 \end{bmatrix} \tag{17}$$

In equations (14)-(16), two hyperparameters α and β are also as inputs to the established Bayesian model. We use the grid search technique based on cross-validation to determine these two hyperparameters. This is an effective tuning method to find an optimal setting. Briefly speaking we can train a learning model using a wide range of values of hyperparameters, evaluate their performance on a hold-out validation set, and select the value that produces the best performance. Another common method for estimating hyperparameter pair α and β is to use an optimization technique that was originally proposed by MacKay [11] and computationally implemented by Foresee & Hagan using Gaussian-Newton algorithm to the Hessian matrix of the error function through an iterative procedure.

3 Experimental results and discussions

In this section, a few numerical examples are presented for the purpose of the test. The first test example is based on a widely studied problem in the machine learning paradigm, called the 'SinC' function problem. The 'SinC' function is the zero-order spherical Bessel function.

A set of 100 data points of $J_0(x)$ is sampled in the interval [-10,10], where x'_i are uniformly distributed and the corresponding y'_i are added in a zero-mean Gaussian noise. To test the robustness of the proposed method to the different level of noises, two level Gaussian noises with the standard deviation 0.1 and 0.2 are experimented respectively. Using a 7-basis functions Bayesian model with hyperparameters $\alpha=10^{-4}$, $\beta=100$ and $\alpha=10^{-4}$, $\beta=25$, two separate regressions have been performed for the above sample data sets. The regression results along with the original data are shown in Figs. 1a&b, where the red solid line is the exact $J_0(x)$ function, the black dashed line denotes the regression curve that connects the means computed from the predictive distribution, and the light shaded region crosses one standard deviation each side of the mean. The experimental results show that the Bayesian model presents a smooth fitting for both level of noises. In the case of lower noise, the means of the predictive distribution well approximate the noisy data as illustrated in Fig1a. For the case of higher noise level as shown in Fig 1b, the overall shape of the predictive curve appears acceptable but some segments deviate from the exact function with a relatively large error, which reflects

the possible enlargement of perturbation from some data points due to large random noises.

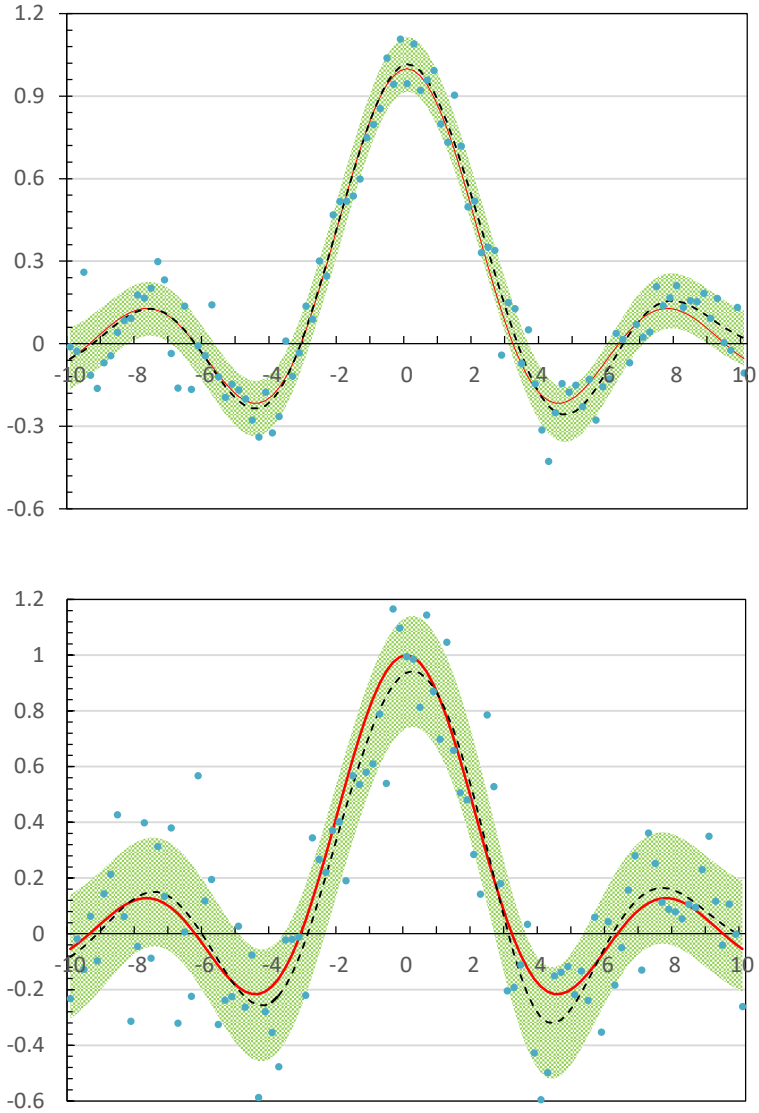

Fig. 1a&b. Bayesian regression in the SinC synthetic data set with 100 points. (a) The standard deviation of added noise is 0.1; and (b) The standard deviation of added noise is 0.2. The red solid line denotes the exact SinC function, while the black dashed line represents the mean from Bayesian model.

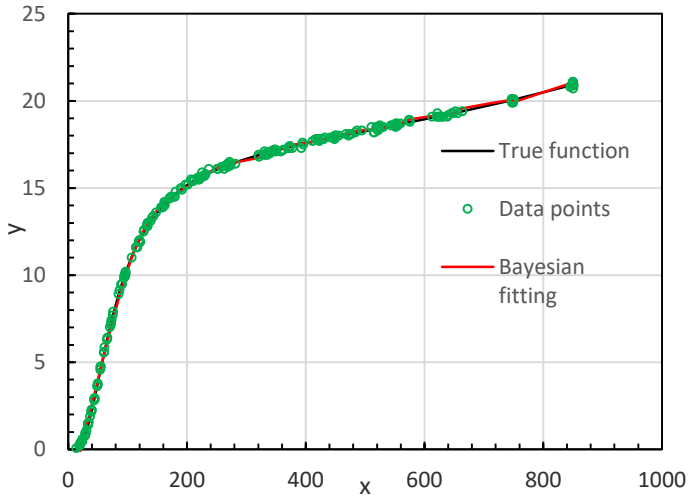

Fig. 2. Fitting result using Bayesian model for Hahn1 function

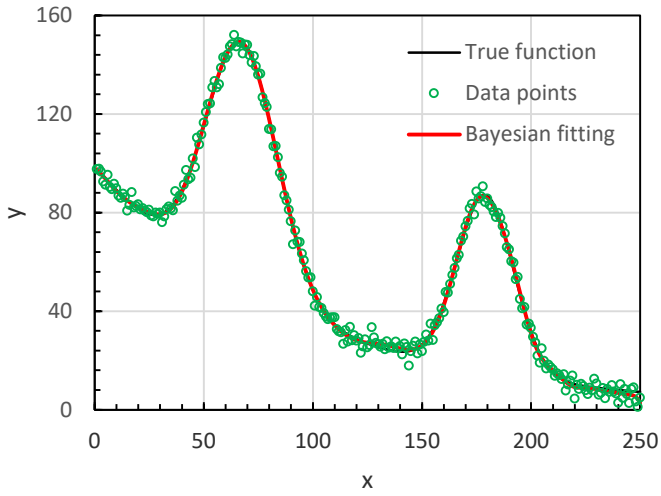

Fig. 3. Fitting result using Bayesian model for Gauss1 dataset

Another two functions we have tested are from the benchmark dataset of the National Institute of Standard Technology (NIST) [12]. They are Hahn1 function and Gauss1 dataset, which are often used to verify the accuracy of new developed nonlinear regression software package. They are generated from the true functions with adding

zero-mean normal distribution noises. The Hahn1 is a 7-parameters rational function and the Gauss1 dataset is generated from two well-separated Gaussians on a decaying exponential baseline plus normally distributed noise. The former can efficiently model the thermal expansion of electrons in copper while the latter is an important category of spectroscopic line shape profile. Figures 2 and 3 shows our Bayesian model well fit these two data sets.

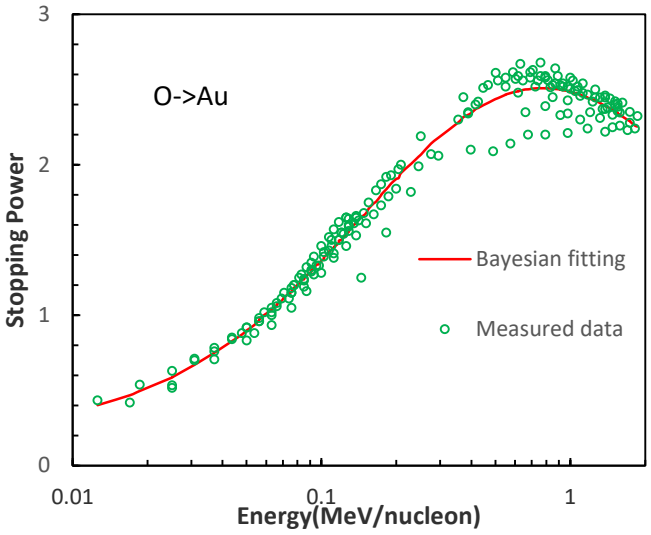

Fig. 4. Fitting results using Bayesian model for stopping power data

There are many practical applications in applied science and engineering using fitting curve method to fit experimentally measured data into curves. As examples to demonstrate our method, we consider a group of data from high energy physics experiments which are related to fitting stopping power curves. We select MeV oxygen projectiles in the target materials including C, Si and Au. The data to be fitted are primarily from the atomic and nuclear data compilations published by Nuclear Data Services, International Atomic Energy Agency (IAEA) (https://www-nds.iaea.org/stopping). From Fig.4, it can be observed that the fitting curve produced by the proposed Bayesian method fits the data points exceptionally well and there is no overfitting issue at all. In addition, the fitting curve reveals the typical features of data point distribution such as peaks. Due to the constraint of pages, only one figure of stopping power data fitting is selected to show.

4 Conclusions

This study presents a theoretical framework based on Bayesian probabilistic model for solving nonlinear curve fitting. With the introduction of an extra linear term, the proposed model enhances the performance of fitting accuracy, offering a new

alternative approach to conventional numerical analysis based method. Conceptually, a better approximation has achieved largely through the hybrid of Gaussian basis functions and a linear function. Relative to the ordinary linear regression model, the proposed method effectively refines the basic regression model with a dual correction – a linear term contribution, and Bayesian posterior information feedback which controls the possible overfitting. Future work will explore the use of the developed method to establish empirical formula based on analysis of curve fitting result. In addition, a further investigation and more tests from diverse datasets should allow the implementation of this method as a proprietary software module to be embedded into a practical intelligent data analysis package for various applications. One limitations of the proposed method is to tune hyperparameters. The Bayesian approach helps to prevent overfitting by controlling model capacity but it gives rise a new issue that a careful tuning for hyperparameters is required. The optimization of hyperparameter itself is a fairly difficult problem. Although several algorithms such as grid search, random search and Bayesian optimization etc. have been developed for applications, smarter tuning methods like random forest algorithm are still being investigated extensively. Another limitation lies at the fixed basis function where the number of basis function actually may need to grow with the dimensionality of the input space in certain circumstance. This problem could be eased if the intrinsic dimensionality of the real data sets is not large due to some correlations of input variables.

References

1. Denison, D. G. T., Mallick, B. K., Smith, A. F. M.: Automatic Bayesian curve fitting. J. R. Statist. Soc. B60, Part 2, 333-350 (1998)
2. Dimatteo, I., Genovese, C. R., Kass, R. E.: Bayesian curve fitting with free-knot Splines. Biometrika 88, 1055-1071 (2001)
3. Green, P. J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model Determination. Biometrika 82, 711-732 (1995)
4. Chen, C., Yu. K.: Automatic Bayesian quantile regression curve fitting. Statistics and Computing 19, 271-281 (2009)
5. Zhou, S., Shen, X.: Spatially adaptive regression splines and accurate knot selection Scheme. Journal of American Statistical Association 96, 247-259 (2001)
6. Koenker, Q.: Quantile Regression. Cambridge University Press, Cambridge (2005)
7. Hansen, M. H., Kooperberg, C.: Spline adaptation in extended linear models. Statistical Science 17, 2-51 (2002)
8. Paul, H.: The stopping power of matter for positive ions. In G. Natanasabapathi (Ed.), Modern Practices in Radiation Therapy (2012)
9. Bishop, C. M.: Pattern recognition and machine learning. Springer (2006)
10. Li, M. M., Verma, B.: Nonlinear curve fitting to stopping power data using RBF neural networks. Expert Systems with Applications 45, 161-171 (2016)
11. Mackay, D. J. C.: Bayesian interpolation. Neural Computation 4, 415−447 (1992)
12. www.itl.nist.gov/div898/strd/nls/da