International Journal for Uncertainty Quantification, 6 (3): 195–213 (2016)

BAYESIAN NONPARAMETRIC GENERAL REGRESSION

Ka-Veng Yuen & Gilberto A. Ortiz***

Faculty of Science and Technology, University of Macau, 999078, Macao, China

*Address all correspondence to: Ka-Veng Yuen, E-mail: kvyuen@umac.mo

Original Manuscript Submitted: 12/15/2015; Final Draft Received: 7/27/2016

Bayesian identification has attracted considerable interest in various research areas for the determination of the mathematical model with suitable complexity based on input-output measurements. Regression analysis is an important tool in which Bayesian inference and Bayesian model selection have been applied. However, it has been noted that there is a subjectivity problem of model selection results due to the assignment of the prior distribution of the regression coefficients. Since regression coefficients are not physical parameters, assignment of their prior distribution is nontrivial. To resolve this problem, we propose a novel nonparametric regression method using Bayesian model selection in conjunction with general regression. In order to achieve this goal, we also reformulate the general regression under the Bayesian framework. There are two attractive features of the proposed method. First, it eliminates the subjectivity of model selection results due to the prior distribution of the regression coefficients. Second, the number of model candidates is drastically reduced, compared with traditional regression using the same number of design/input variables. Therefore, this allows for the consideration of a much larger number of potential design variables. The proposed method will be assessed and validated through two simulated examples and two real applications.

KEY WORDS: *Bayesian inference, general regression, input-output relationship, model selection, nonparametric modeling*

1. INTRODUCTION

The task of mathematical modeling can be complex and it requires in-depth understanding of the physical phenomenon. In particular, nonparametric regression is different from traditional parametric approaches using a prescribed parametric model with unknown parameters to be determined using input-output data. It adjusts the flexible regression surface adaptively according to the data so the prescription of a parametric model (e.g., functional form) is avoided. A number of techniques have been proposed for nonparametric regression and applications can be found in a number of research areas, such as econometrics [1], air quality prediction [2], image processing [3], traffic engineering [4], target detection [5], and insurance risk analysis [6], etc.

Among nonparametric methods, the general regression neural network [7] is a learning algorithm based on the Nadaraya-Watson kernel regression algorithm [8, 9]. In contrast to the back-propagation-based neural networks which require iterations to converge to the optimal model, the general regression neural network involves an efficient one-pass learning step. The general regression neural network is based on statistical principles and it converges to the true regression surface as the number of samples increases. It has been applied successfully to econometrics [10], energy conservation [11], image processing [12], assessment of high power systems [13], and soil dynamics [14], etc. Although the general regression neural network offers a feasible solution to fit a regression surface to the available data, it remains a challenge to select the proper set of design variables (i.e., the inputs to the network). Note that inclusion of more design variables does not necessarily enhance the predictive power of the model.

2152–5080/16/$35.00

Bayesian inference offers a rigorous framework to update the probability distribution of uncertain parameters and to select the most suitable model according to measurements. It has been applied successfully to structural dynamics [15–21], damage detection [22, 23], geotechnical engineering [24–26], air quality prediction [27], climate change [28], random vibrations [29], hydraulic engineering [30], subsurface flow prediction [31], and reliability analysis [32–35], etc. Although general regression is powerful, its formulation in Bayesian is nontrivial. In this paper, we will reformulate the general regression under the Bayesian framework.

On the other hand, Bayesian model selection for traditional regression suffers from the subjectivity of prior distribution assignment. Given some input-output measurement D, the Bayesian model selection approach evaluates the model candidates $C^{(k)}, k = 1, 2, \ldots, N_c$, according to their plausibility $P(C^{(k)}|D)$. By using the Bayes' theorem, the plausibility $P(C^{(k)}|D)$ is proportional to the evidence $p(D|C^{(k)})$, which can be expanded according to the theorem of total probability:

$$p\left(D|C^{(k)}\right) = \int_{\Theta^{(k)}} p\left(D|\boldsymbol{\theta}^{(k)}, C^{(k)}\right) p\left(\boldsymbol{\theta}^{(k)}|C^{(k)}\right) \mathrm{d}\boldsymbol{\theta}^{(k)},\ k = 1, 2, \ldots, N_c, \tag{1}$$

where $\boldsymbol{\theta}^{(k)} \in \Theta^{(k)}$ is the uncertain parameter vector of model $C^{(k)}$. It is clearly seen that the prior distribution has a direct effect on the evidence. For example, consider a sufficiently wide uniform prior distribution that covers all regions with significant likelihood values. If one doubles the width of this prior distribution for one of the parameters, the density of the prior distribution and, hence, the evidence will become half. Note that different regression models are associated with different variables and/or a different number of terms. Therefore, scaling the prior distribution of a parameter affects the evidence of some but not all models. As a result, the model selection result will be arbitrary subject to such scaling. This is different from parametric identification in which the prior distribution will have no effect if it is sufficiently flat.

In this paper, we propose a novel Bayesian approach for nonparametric regression, namely, Bayesian nonparametric general regression. This proposed method does not require prescription of one or more explicit models and it requires only a set of potential design variables to be selected from, without specification of functional form. To achieve this goal, the general regression will be reformulated under the Bayesian framework. There are two attractive features of the proposed method. First, the aforementioned subjectivity in Bayesian model selection for traditional regression problems is resolved. Second, the number of model candidates to be evaluated is drastically smaller than that in traditional regression problems. As a result, with the same computational demand, the proposed method allows for the examination of a much larger number of potential design variables. These will be further elaborated upon in Section 3.4.

The structure of this paper is outlined as follows. Section 2 introduces the fundamentals of general regression. Then, Section 3 presents a novel reformulation of general regression from a Bayesian inference perspective as well as a procedure for the selection of the design variables using Bayesian model selection. Section 4 summarizes the proposed method. Finally, Section 5 presents two simulated examples and two real applications to assess the effectiveness of the proposed method.

2. FUNDAMENTALS OF GENERAL REGRESSION

Consider a dataset $\boldsymbol{D} = \{D_1, \ldots, D_N\}$, where each data point $D_n = \{\mathbf{x}_n, y_n\}$, $n = 1, \ldots, N$, consists of the measured design vector $\mathbf{x}_n \in \mathbb{R}^d$ and the measured quantity of interest $y_n \in \mathbb{R}$. In traditional regression, a mathematical model has to be prescribed for the input-output relationship:

$$y_n = f\left(\mathbf{x}_n; \boldsymbol{\theta}\right) + \varepsilon_n, \tag{2}$$

where $f(\mathbf{x}_n; \boldsymbol{\theta})$ is a prescribed (linear or nonlinear) function governing the relationship between the design variables in $\mathbf{x}$ and the quantity of interest y based on the parameters in $\boldsymbol{\theta}$; and ε_n is the residual (i.e., prediction error) modeled as a Gaussian random variable with zero mean and standard deviation σ_ε. Construction of the functional form $f(.; \boldsymbol{\theta})$ can be a complex process involving in-depth understanding of the physical phenomenon, but it is usually determined in a rather subjective manner for regression problems.

The general regression neural network [7] is a kernel-based regression algorithm which provides estimates of the quantity of interest y based on the design variables in $\mathbf{x}$ without specification of a parametric functional form $f(.;\theta)$. If the joint probability density function (PDF) $p(\mathbf{x}, y)$ of $\mathbf{x}$ and y is known, the prediction of y can be obtained as the conditional mean given $\mathbf{x}$. Specifically, the expected value of y given $\mathbf{x}$ is readily obtained:

$$E[y|\mathbf{x}] = \frac{\int_{-\infty}^{\infty} yp(\mathbf{x}, y)\,\mathrm{d}y}{\int_{-\infty}^{\infty} p(\mathbf{x}, y)\,\mathrm{d}y}, \tag{3}$$

where $E[y|\mathbf{x}]$ is also called the *regression of* y *on* $\mathbf{x}$. Nevertheless, the joint PDF $p(\mathbf{x}, y)$ is usually unknown in practice so a kernel density approximation $\hat{p}(\mathbf{x}, y)$ is used. A popular choice is the Gaussian mixture distribution [36, 37]:

$$\hat{p}(\mathbf{x}, y) = \frac{1}{N(2\pi\sigma_1^2)^{(d+1)/2}} \sum_{n=1}^{N} \left[\exp\left(-\frac{(\mathbf{x}-\mathbf{x}_n)^T(\mathbf{x}-\mathbf{x}_n) + (y-y_n)^2}{2\sigma_1^2} \right) \right]. \tag{4}$$

This estimator converges to $p(\mathbf{x}, y)$ at all points $\{\mathbf{x}_n, y_n\}$, $n = 1, \ldots, N$, asymptotically as N tends to infinity, provided that the *smoothing parameter* $\sigma_1^2 = \sigma_1^2(N)$ satisfies the following criteria [7]:

$$\lim_{N\to\infty} \sigma_1(N) = 0, \tag{5}$$

$$\lim_{N\to\infty} N\sigma_1^d(N) = \infty. \tag{6}$$

Replacing $p(\mathbf{x}, y)$ in Eq. (3) by $\hat{p}(\mathbf{x}, y)$, the conditional mean of y given $\mathbf{x}$ can be approximated as follows:

$$\hat{y}(\mathbf{x}) = E_{\hat{p}}[y|\mathbf{x}; \sigma_1^2] = \frac{\sum_{n=1}^{N} y_n \exp\left[-\left((\mathbf{x}-\mathbf{x}_n)^T(\mathbf{x}-\mathbf{x}_n) \right) / \left(2\sigma_1^2\right) \right]}{\sum_{n=1}^{N} \left[\exp - \left((\mathbf{x}-\mathbf{x}_n)^T(\mathbf{x}-\mathbf{x}_n) \right) / \left(2\sigma_1^2\right) \right]}, \tag{7}$$

where $E_{\hat{p}}[y|\mathbf{x}; \sigma_1^2]$ denotes the mathematical expectation according to the PDF $\hat{p}$. The only unknown parameter in general regression is the smoothing parameter σ_1^2 (sometimes referred to as σ_1). It controls the trade-off between the smoothness of the model output surface and the fitness to the data points. A smooth joint PDF $\hat{p}(\mathbf{x}, y)$ is associated with a large value of σ_1^2 and the model output will become the average of all the measured values of the quantity of interest as $\sigma_1^2 \to \infty$. On the other hand, a small value of σ_1^2 allows the joint PDF to be non-Gaussian but overfitting occurs as $\sigma_1^2 \to 0$. It is intuitive to obtain the value of σ_1^2 by minimizing the mean squared error between the general regression predictions and the corresponding measurements. However, direct application of this approach leads to overfitting because perfect data fitting occurs as $\sigma_1^2 \to 0$. However, the resultant model will give zero prediction except at the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. To resolve this problem, the common approach is to remove the mth element from the summations in Eq. (7) for the prediction of the same point:

$$\hat{y}(\mathbf{x}_m) = \frac{\sum_{\substack{n=1 \\ n\neq m}}^{N} y_n \exp\left[-\left((\mathbf{x}_m-\mathbf{x}_n)^T(\mathbf{x}_m-\mathbf{x}_n) \right) / \left(2\sigma_1^2\right) \right]}{\sum_{\substack{n=1 \\ n\neq m}}^{N} \exp\left[-\left((\mathbf{x}_m-\mathbf{x}_n)^T(\mathbf{x}_m-\mathbf{x}_n) \right) / \left(2\sigma_1^2\right) \right]}. \tag{8}$$

Then, these model predictions will be used to fit the measurements to obtain the optimal value of σ_1^2.

3. PROPOSED BAYESIAN NONPARAMETRIC GENERAL REGRESSION

3.1 Bayesian Formulation of General Regression

The basic concept of general regression was introduced in Section 2. However, Bayesian formulation of general regression is a nontrivial task since the product of the Gaussian distributions of the fitting errors obtained from the

standard general regression formulation in Section 2 does not give the correct likelihood function. In this section, we reformulate the general regression under the Bayesian framework. First, use $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$ to denote the design variable matrix and $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^N$ to denote the measurement vector. Then, C represents the general regression model (i.e., the set of the d design variables in $\mathbf{X}$ included for the prediction of the quantity of interest y) and $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is the associated uncertain parameter vector to be determined. By using the Bayes' theorem, the posterior PDF of $\boldsymbol{\theta}$ is given by [38, 39]

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, C) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}, C)\, p(\boldsymbol{\theta}|C)}{p(\mathbf{y}|\mathbf{X}, C)}, \tag{9}$$

where $p(\mathbf{y}|\mathbf{X}, C)$ is the normalizing constant, $p(\boldsymbol{\theta}|C)$ is the prior PDF of the uncertain parameters to represent the prior knowledge based on user's judgment, and $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}, C)$ is the likelihood function to express the goodness of fit of the data points for given $\boldsymbol{\theta}$.

As stated earlier, it is nontrivial to obtain the likelihood function. In this paper, we expand the likelihood function as the product of conditional PDFs [40]:

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}, C) = p(y_1, y_2, \ldots, y_{N-1}, y_N|\boldsymbol{\theta}, \mathbf{X}, C) = \prod_{m=1}^{N} p(y_m|y_1, y_2, \ldots, y_{m-1}, \boldsymbol{\theta}, \mathbf{X}, C). \tag{10}$$

The conditional PDFs are given by

$$p(y_m|y_1, y_2, \ldots, y_{m-1}, \boldsymbol{\theta}, \mathbf{X}, C) = \left(2\pi\sigma_{2,m}^2\right)^{-1/2} \exp\left[-\frac{\left(y_m - \hat{y}_{m|m-1}(\mathbf{x}_m)\right)^2}{2\sigma_{2,m}^2}\right], \tag{11}$$

where $\hat{y}_{m|m-1}(\mathbf{x}_m) \equiv E[y_m|y_1, y_2, \ldots, y_{m-1}, \boldsymbol{\theta}, \mathbf{X}, C]$ is the regression of y on $\mathbf{X}$ based on the first $m-1$ data points only and it can be computed in a similar way as Eq. (8) using the first $m-1$ data points only:

$$\hat{y}_{m|m-1}(\mathbf{x}_m) = \frac{\sum_{n=1}^{m-1} y_n \exp\left[-\left((\mathbf{x}_m - \mathbf{x}_n)^T(\mathbf{x}_m - \mathbf{x}_n)\right) / \left(2\sigma_{1,m}^2\right)\right]}{\sum_{n=1}^{m-1} \exp\left[-\left((\mathbf{x}_m - \mathbf{x}_n)^T(\mathbf{x}_m - \mathbf{x}_n)\right) / \left(2\sigma_{1,m}^2\right)\right]}. \tag{12}$$

It should be noted that these conditional PDFs contain different data points in the conditioning part so the smoothing parameters $\sigma_{1,m}^2$, $m = 1, 2, \ldots, N$, and the prediction-error variances $\sigma_{2,m}^2$, $m = 1, 2, \ldots, N$, are m-dependent. First, since the smoothing parameters $\sigma_{1,m}^2$, $m = 1, 2, \ldots, N$, represent the sparseness of the data points, they are assumed proportional to the average distance among the data points in the corresponding conditioning part:

$$\sigma_{1,m}^2 = \frac{v_1}{m-1} \sum_{n=1}^{m-1} (\mathbf{x}_m - \mathbf{x}_n)^T (\mathbf{x}_m - \mathbf{x}_n), \tag{13}$$

where v_1 is the *smoothing scale parameter* to be determined.

On the other hand, $\sigma_{2,m}^2$, $m = 1, 2, \ldots, N$, are the prediction-error variances, which depend on the number of data points in the conditioning part and their distance to the point for prediction. By observing Eq. (8), they are assumed to take the following form:

$$\sigma_{2,m}^2 = \frac{v_2}{\sum_{n=1}^{m-1} \exp\left[-2(\mathbf{x}_m - \mathbf{x}_n)^T(\mathbf{x}_m - \mathbf{x}_n)\right]}, \tag{14}$$

where v_2 is the *prediction-error scale parameter* to be determined. For each data point in the conditioning part, it contributes with a positive value to the denominator and, hence, reduces the prediction-error variance. This contribution depends on the distance to the point for prediction. For a smaller distance, the exponential term is larger so the contribution is higher. In this formulation, the smoothing scale parameter v_1 and the prediction-error scale parameter v_2 are the only unknowns to be determined using the data, i.e., $\boldsymbol{\theta} = [v_1, v_2]^T$.

3.2 Parametric Identification of General Regression

In this study, uniform prior distribution is used for the smoothing scale parameter v_1 and the prediction-error scale parameter v_2, i.e., $p(v_1, v_2|C) = (B_{U1} - B_{L1})^{-1}(B_{U2} - B_{L2})^{-1}$ for $v_1 \in (B_{L1}, B_{U1})$, $v_2 \in (B_{L2}, B_{U2})$, and zero otherwise. The support of the prior distribution is assumed to be sufficiently wide so the parametric identification results will depend solely on the data. It will be explained in Section 3.4 and demonstrated in the examples that the width of this prior PDF has no effect on the model selection results in contrast to traditional Bayesian model selection for regression problems.

By considering Eqs. (10)–(14) with uniform prior distribution, the posterior PDF of the unknown parameters can be written as

$$\begin{aligned} p\left(v_1,v_2|\mathbf{y},\mathbf{X},C\right) &\propto p\left(v_1,v_2\right) p\left(\mathbf{y}|v_1,v_2,\mathbf{X},C\right) \\ &\propto \left(v_2\right)^{-(N/2)} \exp\left[-\frac{1}{2v_2}\sum_{m=1}^{N}\Omega_m\left(y_m - \hat{y}_{m|m-1,v_1}\left(\mathbf{x}_m\right)\right)^2\right] \end{aligned} \tag{15}$$

where Ω_m is given by

$$\Omega_m = \sum_{n=1}^{m-1}\exp\left[-2\left(\mathbf{x}_m - \mathbf{x}_n\right)^T\left(\mathbf{x}_m - \mathbf{x}_n\right)\right]. \tag{16}$$

Given v_1, the conditional optimal value for v_2 can be obtained by solving $\partial p(v_1,v_2|\mathbf{y},\mathbf{X},C)/\partial v_2 = 0$:

$$v_2^\star(v_1) = \frac{1}{N}\sum_{m=1}^{N}\Omega_m\left(y_m - \hat{y}_{m|m-1,v_1}(\mathbf{x}_m)\right)^2. \tag{17}$$

However, there is no closed-form solution for the optimal smoothing scale parameter $v_1^\star$ but it can be obtained numerically by maximizing the following function of v_1: $g\left(v_1\right) = p\left(v_1,v_2^\star(v_1)|\mathbf{y},\mathbf{X},C\right)$. By using Eq. (17), the original two-parameter optimization problem is reduced to a one-parameter problem. For any value of v_1, the predictions $\hat{y}_{m|m-1}(\mathbf{x}_m)$ can be estimated using Eq. (12). Then, the conditional optimal prediction-error scale parameter $v_2^\star(v_1)$ can be computed through Eq. (17). Finally, the objective function value $g\left(v_1\right)$ is readily obtained. Standard optimization algorithms can be used to search for the optimal value $v_1^\star$ such that $g\left(v_1\right)$ is maximized, or equivalently, $-\ln g\left(v_1\right)$ is minimized. Meanwhile, the optimal value $v_2^\star$ is equal to the associated conditional optimal $v_2^\star(v_1^\star)$.

3.3 Model Selection and Evidence Evaluation

The general regression approach does not require the specification of a parametric model but the input-output relationship will be obtained adaptively from the data. However, it is still necessary to determine which design variables to include. In this context, a model is referred to a set of design variables x_j, $j = 1, \ldots, d$, for the prediction of the quantity of interest y. Bayesian model selection is used to select the most suitable one among N_c model candidates: $C^{(1)}, C^{(2)}, \ldots, C^{(N_c)}$. In particular, the plausibility of a model can be expressed using the Bayes' rule:

$$P\left(C^{(k)}|\mathbf{y},\mathbf{X}\right) = \frac{p\left(\mathbf{y}|\mathbf{X},C^{(k)}\right)P\left(C^{(k)}\right)}{\sum_{k=1}^{N_c}p\left(\mathbf{y}|\mathbf{X},C^{(k)}\right)P\left(C^{(k)}\right)}, \quad k = 1, 2, \ldots, N_c, \tag{18}$$

where $P(C^{(k)})$ is the prior plausibility on the model $C^{(k)}$ and it is taken to be the same for all models, i.e., $P(C^{(k)}) = 1/N_c$, $k = 1, \ldots, N_c$. On the other hand, the denominator is identical for all models. As a result, the model selection results will depend solely on the data through the evidence $p(\mathbf{y}|\mathbf{X},C^{(k)})$, which plays the vital role in model selection. From the theorem of total probability, the evidence can be expanded as follows:

$$p\left(\mathbf{y}|\mathbf{X},C^{(k)}\right) = \int_0^\infty\int_0^\infty p\left(\mathbf{y}|v_1,v_2,\mathbf{X},C^{(k)}\right)p\left(v_1,v_2|C^{(k)}\right)\mathrm{d}v_1\mathrm{d}v_2, \quad k = 1, 2, \ldots, N_c. \tag{19}$$

By observing Eq. (15), v_2 can be analytically integrated in Eq. (19). As a result, this double integral is deduced to the following single integral:

$$p\left(\mathbf{y}|\mathbf{X},C^{(k)}\right)=\frac{2\Gamma\left((N/2)+1\right)\sqrt{\prod_{m=1}^{N}\Omega_m}}{\left(B_{U1}-B_{L1}\right)\left(B_{U2}-B_{L2}\right)\pi^{N/2}}\int_0^{\infty}\left[\sum_{m=1}^{N}\Omega_m\left(y_m-\hat{y}_{m|m-1,v_1}(\mathbf{x}_m)\right)^2\right]^{-((N/2)+1)}\mathrm{d}v_1, \quad (20)$$

where $\Gamma(\cdot)$ is the Gamma function; B_{U1}, B_{U2}, B_{L1} and B_{L2} are the upper and lower bounds of the prior PDF of the two scale parameters, respectively. By using the Laplace asymptotic approximation [41, 42], this integral can be approximated and the evidence can be computed efficiently as follows:

$$\begin{aligned}p\left(\mathbf{y}|\mathbf{X},C^{(k)}\right)\approx&\frac{2\Gamma\left((N/2)+1\right)\sqrt{2\pi\prod_{m=1}^{N}\left(\Omega_m/|\mathcal{H}_k\left(v_1^{\star}\right)|\right)}}{\left(B_{U1}-B_{L1}\right)\left(B_{U2}-B_{L2}\right)\pi^{N/2}}\\&\times\left[\sum_{m=1}^{N}\Omega_m\left(y_m-\hat{y}_{m|m-1,v_1^{*}}\left(\mathbf{x}_m\right)\right)^2\right]^{-((N/2)+1)},\end{aligned} \quad (21)$$

where $v_1^{\star}$ is the optimal/updated value of the smoothing scale parameter obtained using the method presented in Section 3.2 and $\mathcal{H}_k\left(v_1^{\star}\right)$ is the Hessian of the negative logarithm of the integrand in Eq. (20). In the present case, the Hessian is reduced to a scalar which is the second derivative of the negative logarithm of the integrand in Eq. (20) evaluated at the optimal point. This can be computed efficiently using the finite difference method.

3.4 Advantages of the Proposed Method

There are two major advantages of the proposed Bayesian nonparametric general regression method:

1. From Eq. (21), the choice of the prior distribution affects the evidence only through $\left(B_{U1}-B_{L1}\right)\left(B_{U2}-B_{L2}\right)$ but this effect is identical for all models. According to Eq. (18), the model plausibility is obtained by normalizing the evidence to have total plausibility equal to unity. As a result, the choice of prior distribution will have no effect on the model plausibility and, hence, the model ranking.

2. In contrast to traditional regression, the proposed method does not require one to specify or compare different functional forms for the same set of design variables. For example, consider one design variable x to predict the quantity of interest y. In traditional model selection for the regression model, it is necessary to specify and compare a number of functional forms f of the relationship: $y=f(x)+\varepsilon$, e.g., linear, quadratic, or cosine, etc. However, the functional form will be adaptively and automatically obtained in the proposed method without selection among different candidates. This becomes critically important when there are more design variables to select from. For instance, consider four potential design variables. For traditional regression, there can be 20 or even more possible terms (including various combinations of cross terms). For 20 terms, there will be $2^{20}-1\approx 10^6$ model candidates to be assessed. However, the proposed method will have to assess only $2^4-1=15$ model candidates for four potential design variables. This allows for the evaluation of a far larger set of potential design variables.

3.5 Comparison with Other Bayesian Nonparametric Approaches

The proposed methodology takes full advantage of the predictive power of general regression neural networks and the ability of Bayesian inference for the computation of the smoothing parameter and the selection of the appropriate set of design variables. Moreover, the plausibility obtained in the proposed methodology is computed directly from the evidence using the available data.

On the other hand, other methods, such as Bayesian kernel methods [43], perform variable selection by either:

- Computation of the marginal statistics for each variable or sets of variables based on their effect size. In this case, the design variables are ranked according to a measure of plausibility computed in a rather subjective manner, since hyperparameters need to be provided so as to get a metric of plausibility.
- Definition of weights associated to the design variables. These weights are optimized and/or computed through MCMC. Variables associated with weights near zero are discarded, since they are considered irrelevant in predicting the response variable.

Furthermore, the proposed method solves the subjectivity problem associated with the prior distribution in model selection for regression problems, as stated in Section 3.4. Other nonparametric approaches do not address this problem though.

4. PROCEDURE OF THE PROPOSED ALGORITHM

The proposed Bayesian nonparametric general regression method is summarized as follows:

1. Data normalization: Given a dataset, normalize each of the input and output variables to have zero mean and unity standard deviation [44].
2. Evidence evaluation:

   for $k = 1 : N_c$

   2.1 Compute the optimal parameters $v_1^\star$ and $v_2^\star$:
   - Minimize $-\ln g\left(v_1\right) = -\ln p\left(v_1, v_2^\star\left(v_1\right) | \mathbf{y}, \mathbf{X}, C^{(k)}\right)$ with respect to v_1 to obtain $v_1^\star$.
     For given v_1, compute $\hat{y}_{m|m-1}(\mathbf{x}_m)$ using Eq. (12).
     Compute $v_2^\star\left(v_1\right)$ using Eq. (17).
   - $v_2^\star = v_2^\star\left(v_1^\star\right)$

   2.2 Compute the evidence of $C^{(k)}$ using Eq. (21).

   end for
3. Compute the plausibility of each model using Eq. (18).
4. Select the model with the highest plausibility.

Then, the trained model can be used for prediction using Eq. (7) with the smoothing parameter σ_1^2 given by

$$\sigma_1^2 = \frac{v_1^\star}{N} \sum_{n=1}^{N} \left(\mathbf{x} - \mathbf{x}_n\right)^T \left(\mathbf{x} - \mathbf{x}_n\right), \tag{22}$$

where $\mathbf{x}$ is the design vector of the point for prediction. Finally, the variance of the prediction error can be estimated in a similar fashion to Eq. (14) as follows:

$$\sigma_2^2 = \frac{v_2^\star}{\sum_{n=1}^{N} \exp\left[-2\left(\mathbf{x} - \mathbf{x}_n\right)^T \left(\mathbf{x} - \mathbf{x}_n\right)\right]}. \tag{23}$$

It should be pointed out that extrapolation, i.e., prediction of points outside the region where the training data are located, is not recommended. From Eq. (7) it can be observed that the estimations computed via general regression are a nonlinear weighted average of the training data. In this case, prediction of points outside the training data region will tend to be erroneous. Moreover, those predictions will be associated with a high variance value, since the variance now depends on the distance from the samples to the point of prediction [see Eq. (23)]. Given these reasons, extrapolation should be avoided in general.

5. ILLUSTRATIVE EXAMPLES

5.1 Example 1: Simulated Data

In the first example, the following input-output relationship is considered:

$$\begin{aligned} y_n = \exp\left(-0.05x_{1,n}\right)\left[0.5 - 0.1x_{1,n} - 0.9x_{1,n}^2 + 3x_{1,n}^4 + 0.5\left|x_{1,n}\right|\sin\left(10x_{1,n}\right) + 0.4\sin\left(5x_{1,n}^4\right)\right. \\ \left. - 0.2\sin\left(-2 - x_{1,n}^3\right) - 2.25\exp\left(-0.1x_{1,n}^2\right) - 0.5\exp\left(-0.1 - 1.2x_{1,n}^3\right)\right] + \varepsilon_n. \end{aligned} \tag{24}$$

The dataset contains $N = 300$ data points, in which $x_{1,n}$, $n = 1, \ldots, N$, were generated according to the uniform distribution over $(-0.8,\, 0.9)$; ε_n, $n = 1, \ldots, N$, were generated from the Gaussian distribution with zero mean and standard deviation σ_ε taken as 10% of the standard deviation of the actual quantity of interest. The relationship in Eq. (24) and the parameters were used for data generation purpose only. In other words, the complicated functional form in Eq. (24) and the parameter values were assumed unknown in the entire identification process.

In addition to x_1, two extra variables x_2 and x_3 were generated according to the following autoregressive processes:

$$x_{2,n} = 0.05 + 0.6x_{2,n-1} + 0.2x_{2,n-2} - 0.1x_{2,n-3} + \varepsilon_{2,n}, \tag{25}$$

$$x_{3,n} = 0.01 + 0.7x_{3,n-1} + 0.4x_{3,n-2} - 0.1x_{3,n-3} - 0.2x_{3,n-4} + \varepsilon_{3,n}, \tag{26}$$

where $\varepsilon_{2,n}$ and $\varepsilon_{3,n}$, $n = 1, 2, \ldots, N$, are independent standard Gaussian random variables. The autoregressive models in Eqs. (25) and (26) and the parameter values are considered unknown in the identification process and they are used for data generation purpose only.

There are three potential design variables, namely, x_1, x_2, and x_3, for the prediction of the quantity of interest y. By considering different combinations, seven model candidates will be evaluated:

- $C^{(1)} : x_1$
- $C^{(2)} : x_2$
- $C^{(3)} : x_3$
- $C^{(4)} : x_1, x_2$
- $C^{(5)} : x_1, x_3$
- $C^{(6)} : x_2, x_3$
- $C^{(7)} : x_1, x_2, x_3$

Note that there will be far more model candidates for evaluation in traditional regression because there are many functional forms to be compared even for a given set of design variables.

In order to perform model selection from the seven model candidates and assess the obtained model, the original dataset was partitioned into a training dataset with 70% randomly selected data ($N_{tr} = 210$ points) and a testing dataset with the remaining 30% of the data ($N_{te} = 90$ points). It is expected the model $C^{(1)}$ will be selected since x_1 is the only variable involved in the actual input-output relationship in Eq. (24).

Table 1 shows the identification and model selection results using the training dataset with the uniform prior distribution $U(0, 50)$ for both parameters. The models are listed in descending order of plausibility, i.e., from the most plausible to the least plausible. The first column lists the set of variables in each model. For example, "$(1\,2)$" denotes the model with design variables x_1 and x_2, i.e., $C^{(4)}$. The second and third columns present the updated smoothing scale parameter $v_1^\star$ and the updated prediction-error scale parameter $v_2^\star$, respectively. The fourth and fifth columns show the maximum likelihood value and evidence for each model, respectively. Finally, the last column presents the model plausibility. Model "(1)" achieves the highest maximum likelihood value, even over other models with additional design variables. It may be counterintuitive for the simplest model to fit the data the best. This is because the more complicated models are not associated with more adjustable parameters so they do not possess higher

TABLE 1: Model selection results (Example 1)

Model	$v_1^\star$	$v_2^\star$	$p\left(\mathbf{y}\middle\vert v_1^\star, v_2^\star, \mathbf{X}, C^{(k)}\right)$	$p\left(\mathbf{y}\middle\vert \mathbf{X}, C^{(k)}\right)$	$P\left(C^{(k)}\middle\vert \mathbf{y}, \mathbf{X}\right)$
(1)	0.0015	0.5047	4.31×10^{45}	7.72×10^{38}	1.00
$(1\,2)$	0.0060	0.6594	3.61×10^{-25}	7.36×10^{-32}	9.56×10^{-71}
$(1\,3)$	0.0038	0.6999	5.36×10^{-26}	1.47×10^{-32}	1.90×10^{-71}
$(1\,2\,3)$	0.0110	0.6035	7.48×10^{-81}	3.86×10^{-87}	5.01×10^{-126}
(2)	5.59×10^{-12}	35.99	3.72×10^{-149}	1.70×10^{-146}	2.21×10^{-185}
(3)	1.0968	33.10	2.19×10^{-148}	3.18×10^{-153}	4.12×10^{-192}
$(2\,3)$	4.73×10^{-7}	11.42	1.20×10^{-154}	4.72×10^{-164}	6.10×10^{-203}

flexibility in data fitting. More importantly, by introducing additional but unrelated design variables, the distance (metric) structure of the data points is distorted. The underlying concept of general regression is to assign higher weightings for the data points closer to the point for prediction. However, the additional unrelated design variables impose undesirable reordering of the distance ranking and, hence, inappropriate assignment on the weightings. As a result, model "(1)" is preferred over all other models with plausibility virtually equal to unity. This is consistent with the parsimony principle that the simplest model explaining the data should be selected. In this case, the simplest model "(1)" fits the data the best so it is selected with an overwhelmingly high plausibility. This result is reconfirmed by the fact that x_1 is the only design variable in the actual input-output relationship. It highlights the success of the proposed methodology for this highly complicated unknown input-output relationship.

Then, the testing dataset is used to assess the predictability of the obtained model. This dataset was not involved in the model training process so it can be employed to examine the model generalization capability for new data. Figure 1 demonstrates the prediction results of the obtained model using the testing dataset. The true input-output relationship, the testing data points, and the model prediction are represented by the solid line, solid circles, and dashed line, respectively. By using the proposed method, not only the prediction but also its associated uncertainty

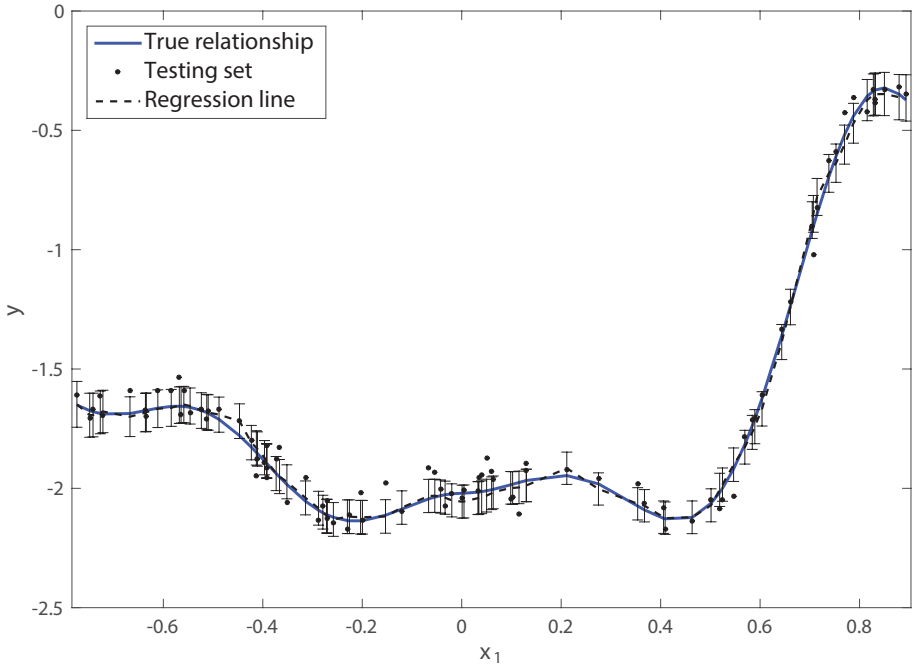

FIG. 1: True relationship and estimated model using testing dataset (Example 1)

can be estimated. In this figure, error bars are provided to cover the 90% credibility interval for each prediction. The predicted curve is in good agreement with the true input-output relationship, which reconfirms the predictive power of the proposed Bayesian nonparametric regression methodology.

It is worth noting that the estimated parameters $v_1^\star$ and $v_2^\star$ are maximum a posteriori values of the smoothing scale parameter and the prediction-error variance term. These parameters provide minimum error and minimum variance of the predictions. Any other set of values will provide predictions with a larger variance, since erroneous predictions are bound to happen due to overfitting (in case of a small smoothing parameter), or due to a poor fitting caused by a large smoothing parameter. By observing Eqs. (17) and (23), larger prediction errors will inevitably enlarge the prediction-error variance of the predictions. As an example, Fig. 2 shows the prediction results in the same fashion as Fig. 1. In this case, the smoothing parameter v_1 was set to 10% of the value of $v_1^\star$. From Fig. 2 it can be observed that the regression line is not as smooth as the one in Fig. 1. Moreover, the prediction-error scale parameter v_2 was 32% higher compared to $v_2^\star$ in Table 1. The same behavior with respect to the variance is observed if the smoothing scale parameter is set to a higher value than $v_1^\star$.

We have explained in Section 1 the subjectivity of Bayesian model selection for traditional regression due to the prior distribution even when it is sufficiently flat. To reconfirm that this problem is resolved by the proposed method, we rerun model selection using different uniform distributions, namely, $U(0,40)$ and $U(0,100)$, and the results are shown in Table 2. First, the model rankings are identical to Table 1. As explained previously in Section 3.4, it is not surprising to observe the difference of evidence $p(\mathbf{y}|\mathbf{X},C^{(k)})$ among the three prior distributions. However, as expected, the plausibility results are identical for all three prior distributions. Model "(1)" is still overwhelmingly preferred over the others and this reconfirms that the width of the prior PDF has no effect on model selection results in the proposed method.

5.2 Example 2: Simulated Data

In this example, the following input-output relationship is considered:

$$y_n = 2 + 3\sin(x_{1,n}) + 2\exp(x_{2,n}) - 3\cos(x_{1,n}x_{2,n}) - \sin\left(x_{1,n}^2\right) + x_{2,n}^2 - \exp(x_{1,n}x_{2,n}) + \varepsilon_n \tag{27}$$

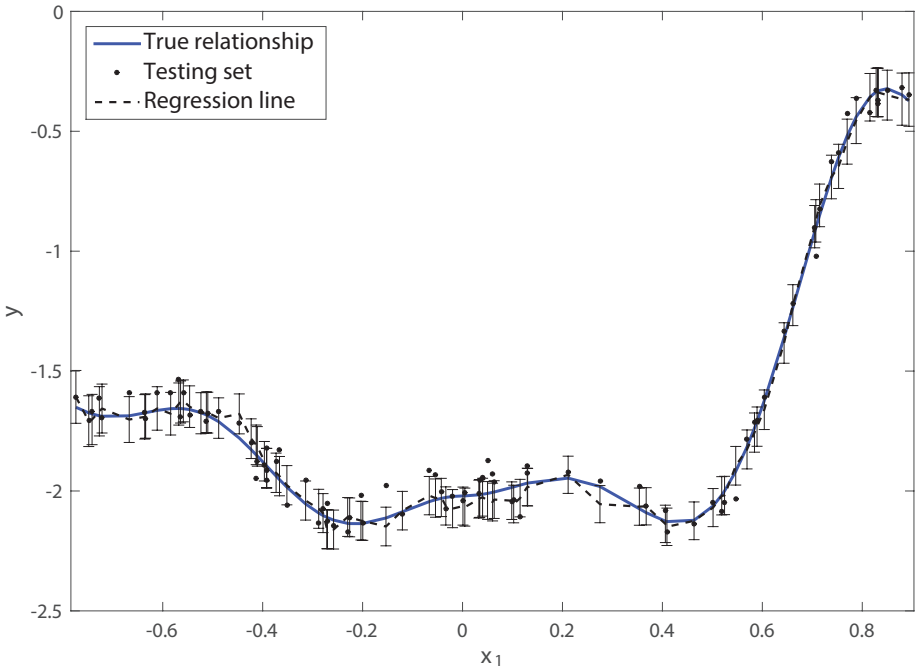

FIG. 2: True relationship and estimated model using testing dataset and smoothing parameter $v_1 = 0.1v_1^\star$ (Example 1)

TABLE 2: Model selection results for two other prior PDFs (Example 1)

$U(0, 40)$			$U(0, 100)$		
Model	$p(\mathrm{y}\|\mathrm{X}, C^{(k)})$	$P(C^{(k)}\|\mathrm{y}, \mathrm{X})$	**Model**	$p(\mathrm{y}\|\mathrm{X}, C^{(k)})$	$P(C^{(k)}\|\mathrm{y}, \mathrm{X})$
(1)	1.20×10^{39}	1.00	(1)	1.93×10^{38}	1.00
(1 2)	1.15×10^{-31}	9.56×10^{-71}	(1 2)	1.84×10^{-32}	9.56×10^{-71}
(1 3)	2.29×10^{-32}	1.90×10^{-71}	(1 3)	3.67×10^{-33}	1.90×10^{-71}
(1 2 3)	6.03×10^{-87}	5.01×10^{-126}	(1 2 3)	9.64×10^{-88}	5.01×10^{-126}
(3)	2.66×10^{-146}	2.21×10^{-185}	(3)	4.26×10^{-147}	2.21×10^{-185}
(2)	4.96×10^{-153}	4.12×10^{-192}	(2)	7.94×10^{-154}	4.12×10^{-192}
(2 3)	7.35×10^{-164}	6.10×10^{-203}	(2 3)	1.18×10^{-164}	6.10×10^{-203}

There are $N = 300$ data points in the dataset, in which $x_{1,n}$ and $x_{2,n}$, $n = 1, \ldots, N$, were generated from the uniform distribution over $(0, 1)$; ε_n, $n = 1, \ldots, N$, were generated from the Gaussian distribution with zero mean and standard deviation σ_ε equal to 20% of the standard deviation of the actual quantity of interest. Note that the data in this example is substantially more polluted than the previous example. Again, the relationship in Eq. (27) and the parameters were used for data generation purpose only. In other words, the functional form and the parameter values were assumed unknown in the identification process.

In this example, four potential design variables, x_j, $j = 1, \ldots, 4$, are considered. The variables x_1 and x_2 are the ones appearing in Eq. (27). Then, two additional variables x_3 and x_4 were generated according to the autoregressive models in Eqs. (25) and (26), respectively. By considering different combinations of these four potential design variables, there are 15 model candidates in this example. It is anticipated that the model with the design variables x_1 and x_2 will be selected.

In order to perform model selection and assess the obtained model, the original dataset was split into a training dataset with 70% randomly selected data ($N_{tr} = 210$ points) and a testing dataset with the remaining 30% of the data ($N_{te} = 90$ points). Table 3 shows the identification and model selection results using the training dataset and prior

TABLE 3: Model selection results (Example 2)

Model	$v_1^\star$	$v_2^\star$	$p(\mathrm{y}\|v_1^\star v_2^\star, \mathrm{X}, C^{(k)})$	$p(\mathrm{y}\|\mathrm{X}, C^{(k)})$	$P(C^{(k)}\|\mathrm{y}, \mathrm{X})$
(1 2)	0.0249	0.6047	8.68×10^{-19}	1.57×10^{-24}	1.00
(1 2 4)	0.0270	0.2476	2.98×10^{-38}	1.13×10^{-43}	7.18×10^{-20}
(1 2 3)	0.0298	0.2620	4.71×10^{-41}	1.59×10^{-46}	1.01×10^{-22}
(2)	0.0285	4.8881	1.04×10^{-59}	5.92×10^{-66}	3.77×10^{-42}
(1 2 3 4)	0.0338	0.1048	2.75×10^{-61}	2.09×10^{-66}	1.33×10^{-42}
(2 4)	0.0334	1.8285	2.32×10^{-72}	2.72×10^{-78}	1.73×10^{-54}
(2 3)	0.0430	2.0109	4.52×10^{-78}	4.96×10^{-84}	3.16×10^{-60}
(2 3 4)	0.0419	0.7214	1.23×10^{-90}	3.33×10^{-96}	2.11×10^{-72}
(1)	0.1827	29.2624	3.79×10^{-139}	3.66×10^{-145}	2.32×10^{-121}
(1 4)	0.1146	9.2372	1.03×10^{-143}	1.51×10^{-149}	9.58×10^{-126}
(4)	0.1103	30.6434	2.34×10^{-144}	1.56×10^{-150}	9.93×10^{-127}
(3)	4.0930	32.9602	2.96×10^{-148}	4.24×10^{-152}	2.69×10^{-128}
(1 3)	0.2029	10.0887	1.90×10^{-147}	4.64×10^{-153}	2.95×10^{-129}
(3 4)	0.1876	10.0167	7.68×10^{-150}	2.66×10^{-155}	1.69×10^{-131}
(1 3 4)	0.1376	3.0955	1.92×10^{-152}	8.60×10^{-158}	5.47×10^{-134}

distribution $U(0, 50)$ for both parameters. The models are listed in descending order of plausibility. The first column shows the variables included in each model. For example, "$(2\,3\,4)$" denotes the model involving variables $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$. The second and third columns show the updated smoothing scale parameter $v_1^\star$ and the updated prediction-error scale parameter $v_2^\star$, respectively. Then, the fourth and fifth columns contain the maximum likelihood and evidence of each model, respectively. Finally, the last column shows the plausibility of each model. It can be observed that the model "$(1\,2)$" achieves the highest maximum likelihood value, even over the second and third most plausible models with one extra design variable. As in the previous example, models with more design variables do not necessarily fit the data better because they do not possess more adjustable parameters and the additional unrelated design variables distorted the distance structure. The most plausible model "$(1\,2)$" is selected with an overwhelmingly high plausibility of virtually 100%. Indeed, the actual input-output relationship involves the same set of variables, i.e., x_1 and x_2. This reconfirms the capability of the proposed method in identifying the set of proper design variables.

Next, the testing dataset is used to assess the obtained model. Since the testing dataset was not involved in the model training process, it can be employed to examine the model generalization capability for new data. Figure 3 shows the model prediction ($\hat{y}$) versus the true quantity of interest ($\tilde{y}$), and the 45° line of perfect match is also drawn for reference. Furthermore, the error bars are shown to represent the 90% credibility interval for each prediction. Even though the measurement noise level is rather substantial, the model predictions are in good agreement with the actual quantity of interest. Moreover, it can be seen that the actual quantity of interest falls mostly within the credibility intervals, indicating reasonable estimation of the prediction uncertainty.

It can also be observed in Fig. 3 that the extreme points are the ones being poorly fitted with respect to the other points. This is due to the combined effect of two situations:

1. The number of training points at the extremes was not as big as the total training data points in other sections. This can be evidenced by the credibility intervals, which are bigger for the samples in the extremes than for samples in the middle section.

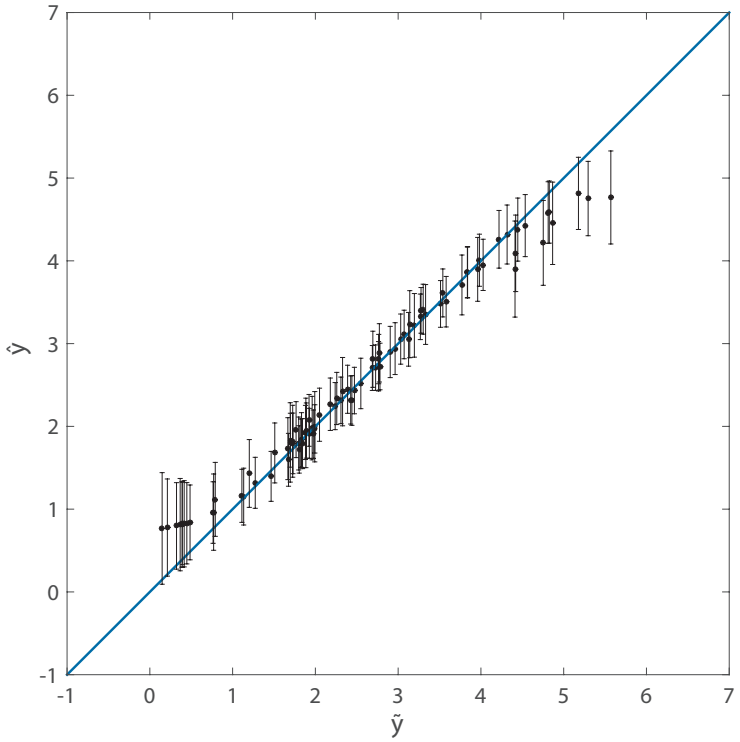

FIG. 3: Model prediction versus actual values of the testing dataset (Example 2)

2. From Eq. (7) it can be observed that the kernel used in GRNN is Gaussian. Samples in the extremes tend to have poor fit due to the symmetry of this kernel, since only points to the right (or left) of these extreme samples provide information for the inference.

Finally, to verify that the model selection results are independent from the prior distribution, the proposed method is rerun for two other prior distributions [namely, $U(0, 40)$ and $U(0, 100)$] and the results are presented in Table 4. As in the previous example, it is not surprising that the evidence is affected by the prior distributions but the model plausibility and model ranking are identical for all three prior distributions [including $U(0, 50)$ shown in Table 3]. The model associated with x_1 and x_2 is selected with the plausibility of virtually 100% regardless of the choice of prior distribution. This reconfirms that the proposed method resolves the subjectivity problem due to the prior distribution.

5.3 Example 3: Stature, Hand Length, and Foot Length Data

Modern sciences (such as anthropology, forensics, and medical sciences) benefit from reliable empirical models for the estimation of body segment proportions using incomplete measurements of certain body parts. For instance, forensic anthropologists are able to determine the stature height of a victim based on a single body part (such as a hand, foot, or even just a piece of large bone) when it is all that remains from the victim.

In a recent study performed at the Medical Faculty of the Cukurova University in Turkey [45], researchers collected the data of hand length (HL), foot length (FL), and stature (y) of 155 students to investigate the relationship among these three variables. The objective is to obtain an empirical model to determine the stature of an individual using the hand length, foot length, or both. These data can be found at [46].

Again, a training dataset with 70% randomly selected data ($N_{tr} = 108$ points) and a testing dataset with the remaining 30% of the data ($N_{te} = 47$ points) were obtained from the original dataset. In this case, there are two design variables (HL and FL) so there are three model candidates for selection:

- $C^{(1)} : HL$
- $C^{(2)} : FL$

TABLE 4: Model selection results for two other prior PDFs (Example 2)

$U(0, 40)$			$U(0, 100)$		
Model	$p(\mathbf{y}\|\mathbf{X}, C^{(k)})$	$P(C^{(k)}\|\mathbf{y}, \mathbf{X})$	**Model**	$p(\mathbf{y}\|\mathbf{X}, C^{(k)})$	$P(C^{(k)}\|\mathbf{y}, \mathbf{X})$
(1 2)	2.46×10^{-24}	1.00	(1 2)	3.93×10^{-25}	1.00
(1 2 4)	1.77×10^{-43}	7.18×10^{-20}	(1 2 4)	2.82×10^{-44}	7.18×10^{-20}
(1 2 3)	2.49×10^{-46}	1.01×10^{-22}	(1 2 3)	3.98×10^{-47}	1.01×10^{-22}
(2)	9.27×10^{-66}	3.77×10^{-42}	(2)	1.48×10^{-66}	3.77×10^{-42}
(1 2 3 4)	3.27×10^{-66}	1.33×10^{-42}	(1 2 3 4)	5.23×10^{-67}	1.33×10^{-42}
(2 4)	4.26×10^{-78}	1.73×10^{-54}	(2 4)	6.81×10^{-79}	1.73×10^{-54}
(2 3)	7.78×10^{-84}	3.16×10^{-60}	(2 3)	1.24×10^{-84}	3.16×10^{-60}
(2 3 4)	5.20×10^{-96}	2.11×10^{-72}	(2 3 4)	8.32×10^{-97}	2.11×10^{-72}
(1)	5.72×10^{-145}	2.32×10^{-121}	(1)	9.15×10^{-146}	2.32×10^{-121}
(1 4)	2.35×10^{-149}	9.58×10^{-126}	(1 4)	3.77×10^{-150}	9.58×10^{-126}
(4)	2.44×10^{-150}	9.93×10^{-127}	(4)	3.91×10^{-151}	9.93×10^{-127}
(3)	6.60×10^{-152}	2.69×10^{-128}	(3)	1.06×10^{-152}	2.69×10^{-128}
(1 3)	7.26×10^{-153}	2.95×10^{-129}	(1 3)	1.16×10^{-153}	2.95×10^{-129}
(3 4)	4.15×10^{-155}	1.69×10^{-131}	(3 4)	6.65×10^{-156}	1.69×10^{-131}
(1 3 4)	1.35×10^{-157}	5.47×10^{-134}	(1 3 4)	2.15×10^{-158}	5.47×10^{-134}

- $C^{(3)} : HL, FL$

Table 5 shows the identification and model selection results in the same fashion as Table 1. These results are associated with uniform prior distribution $U(0, 10)$ for both parameters. The model associated only with foot length [i.e., "(FL)"] is overwhelmingly preferred with plausibility virtually equal to 100%. This result agrees with the conclusion presented in [47,48] that the foot length is sufficient for the prediction of the stature of an individual. It can also be observed that the most plausible model is associated with the highest maximum likelihood value. We will verify with the testing dataset that the selected model indeed provides the best predictions among these three models.

Next, the testing dataset is used to evaluate the model generalization capability for new data. Figure 4 shows the predicted stature ($\hat{y}$) using the selected model versus the observed stature (y) of the 47 individuals in the testing dataset and the 45° line of perfect match is also drawn for reference. Since the actual values of the quantity of interest are not known, the measured values are used in this real application in contrast to the previous examples. Again, the model predictions exhibit good agreement with the real measurements, reconfirming the generalization capability of the model. Furthermore, the error bars are also shown to represent the 90% credibility interval for each prediction and it can be seen that the measured quantity of interest mostly falls in the corresponding credibility interval.

TABLE 5: Model selection results (Example 3)

Model	$v_1^\star$	$v_2^\star$	$p(\mathbf{y}\|v_1^\star, v_2^\star, \mathbf{X}, C^{(k)})$	$p(\mathbf{y}\|\mathbf{X}, C^{(k)})$	$P(C^{(k)}\|\mathbf{y}, \mathbf{X})$
(FL)	0.0078	1.3315	6.78×10^{-26}	1.95×10^{-30}	1.00
$(HL\,FL)$	0.0011	0.6943	1.12×10^{-31}	5.85×10^{-37}	3.00×10^{-7}
(HL)	0.0294	2.7676	1.17×10^{-38}	5.06×10^{-43}	2.60×10^{-13}

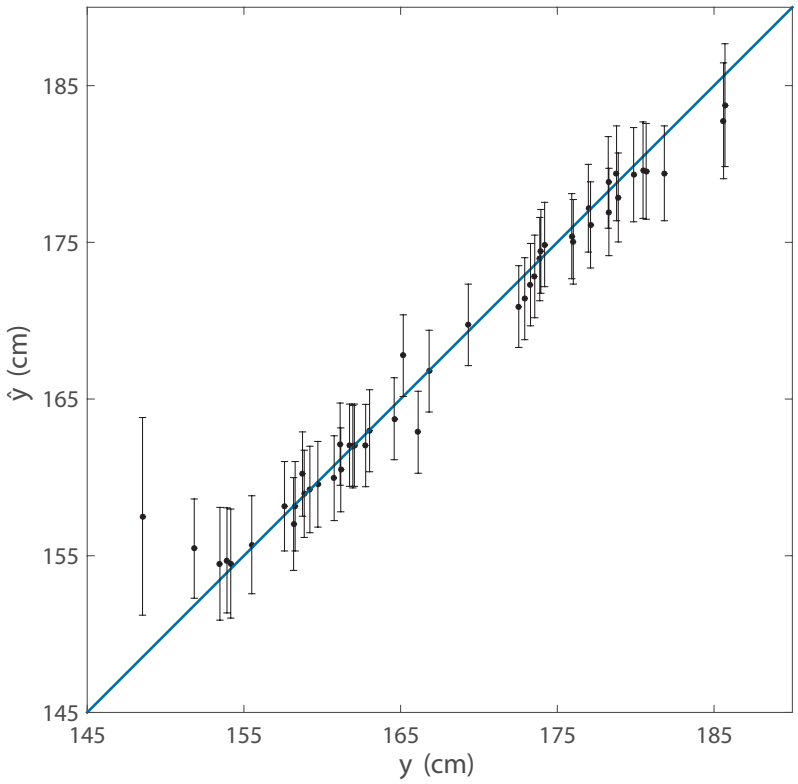

FIG. 4: Model prediction versus measured values of the testing dataset (Example 3)

To verify the model selection result presented in Table 5, the mean-squared error (MSE) of the prediction by all three models is calculated for the testing dataset and presented in Table 6. The selected model (FL) achieves the lowest MSE, even though the testing dataset was not involved in the model training process. The MSE value of the second model is 32% larger than that of the selected model. This highlights the fact that the proposed methodology is able to select the model with best predictive power.

5.4 Example 4: Airfares for US Domestic Routes Estimation

The US Department of Transportation collected data about airfares and passengers for US domestic routes for the fourth quarter of 2002. The dataset used in this example was retrieved from [46] and it comprises the data of the following variables:

- x_1: Distance between two cities
- x_2: Average weekly passengers
- x_3: Market share of the leading airline
- x_4: Average fare of leading airline
- x_5: Market share of low-price airline
- x_6: Average fare of low-price airline
- y: Average fare

The proposed method is used to estimate the average fare (y) using some of the covariates: x_1, x_2, $\ldots$, x_6. By considering different combinations of these six potential design variables, there are 63 model candidates to be selected from. The dataset comprises $N = 200$ points. A training dataset with 70% randomly selected data ($N_{tr} = 140$ points) and a testing dataset with the remaining 30% of the data ($N_{te} = 60$ points) were defined.

Table 7 shows the model selection results using the training dataset. Due to space consideration, only the first ten most plausible models are shown among the 63 model candidates. The first column shows the variables included in the model. For example, "$(2\,3\,4)$" denotes the model involving variables x_2 (average weekly passengers), x_3 (market share of the leading airline), and x_4 (average fare of leading airline). The second and third columns show the updated smoothing scale parameter $v_1^\star$ and the updated prediction-error scale parameter $v_2^\star$, respectively. The fourth and fifth columns contain the maximum likelihood values and evidence of each model. Finally, the last column shows the plausibility of each model. It can be observed that the model associated with the design variables x_4 and x_6 is preferred over the others. From the last column in Table 7, this model accounts for virtually 100% of the plausibility. Meanwhile, the most plausible model is associated with the highest maximum likelihood value.

From the results presented in Table 7, it can be seen that most of the models with high plausibility contain variables x_4 (average fare of leading airline) and x_6 (average fare of low-price airline). This sheds light on the fact that the combined information provided by these two variables is important for the inference of the average fare for domestic flights in the US. In particular, the design variable x_4 (average fare of leading airline) appears in all top ten models. It is not surprising because this variable provides a scale of average fare. On the other hand, the design variable x_3 (market share of the leading airline) provides little information as it appears only in one of the top ten models.

TABLE 6: Mean-squared-error of the testing dataset for each model (Example 3)

Model	MSE
(FL)	326.88
$(HL\,FL)$	432.84
(HL)	1037.50

TABLE 7: Model selection results (Example 4)

Variables	$v_1^\star$	$v_2^\star$	$p\left(\mathbf{y}\middle\vert v_1^\star, v_2^\star, \mathbf{X}, C^{(k)}\right)$	$p\left(\mathbf{y}\middle\vert \mathbf{X}, C^{(k)}\right)$	$P\left(C^{(k)}\middle\vert \mathbf{y}, \mathbf{X}\right)$
$(4\,6)$	0.0100	0.1641	2.62×10^{28}	2.12×10^{24}	1.00
(4)	0.0099	0.4521	6.70×10^{19}	3.42×10^{15}	1.62×10^{-9}
$(2\,4\,6)$	0.0138	0.1102	2.80×10^{17}	4.29×10^{13}	2.03×10^{-11}
$(1\,4\,6)$	0.0072	0.0993	1.68×10^{17}	1.54×10^{13}	7.29×10^{-12}
$(4\,5\,6)$	0.0138	0.0955	1.43×10^{15}	2.78×10^{11}	1.31×10^{-13}
$(3\,4\,6)$	0.0146	0.0990	1.61×10^{8}	3.75×10^{4}	1.77×10^{-20}
$(2\,4)$	0.0142	0.3067	1.31×10^{8}	7.24×10^{3}	3.42×10^{-21}
$(1\,4)$	0.0071	0.2728	4.13×10^{7}	2.48×10^{3}	1.17×10^{-21}
$(2\,4\,5\,6)$	0.0152	0.0610	2.77×10^{5}	7.96×10^{1}	3.76×10^{-23}
$(1\,2\,4\,6)$	0.0167	0.0696	8.66×10^{4}	2.28×10^{1}	1.08×10^{-23}

As in the previous examples, the testing dataset is used to evaluate the generalization capability of the models for new data. Figure 5 shows the predicted airfare for US domestic flights ($\hat{y}$) using the selected model versus the measured airfares (y) and the 45° line of perfect match is also drawn for reference. It can be observed that the proposed methodology successfully recovers the input-output relationship, even though there are two points deviating from the 45° line. It is worth pointing out that the original data comprises different domestic routes and different airlines. Even though it is suspected that the relationship between the design variables and the measured airfare is

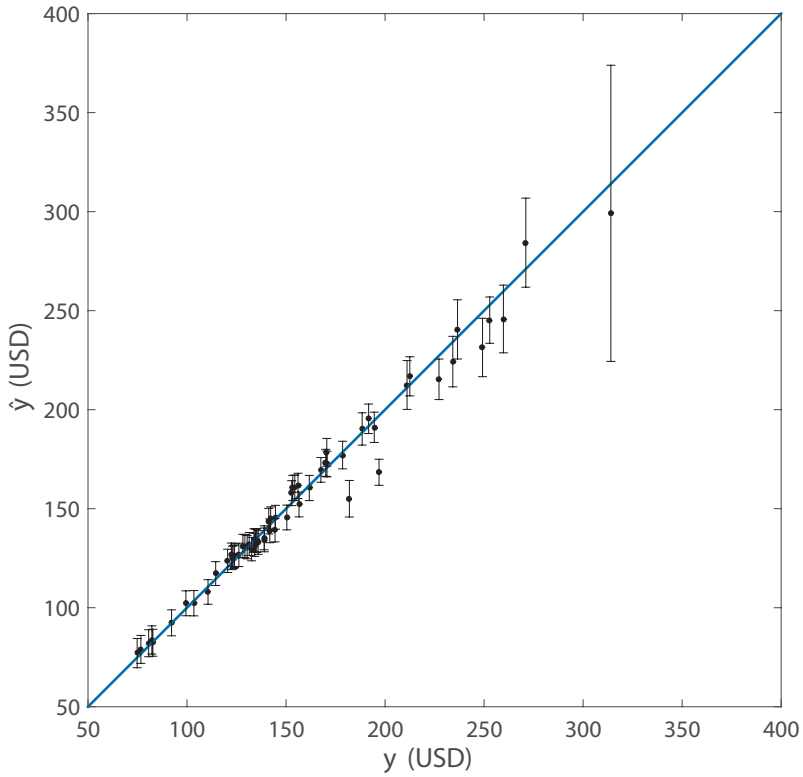

FIG. 5: Model prediction versus measured values using testing dataset (Example 4)

TABLE 8: Mean-squared-error of the testing dataset for each model (Example 4)

Model	MSE
$(4\,6)$	54.92
(4)	56.49
$(2\,4\,6)$	93.65
$(1\,4\,6)$	55.77
$(4\,5\,6)$	59.86
$(3\,4\,6)$	68.66
$(2\,4)$	88.60
$(1\,4)$	96.65
$(2\,4\,5\,6)$	100.0
$(1\,2\,4\,6)$	101.1

region dependent, this result can be considered acceptable. Furthermore, the error bars are also shown to represent the 90% credibility interval for each prediction and it can be seen that the measured quantity of interest falls mostly within the corresponding intervals.

Finally, to verify the suitability of the selected model, the mean-squared error of the predictions by all models with the testing dataset are shown in Table 8. The models are listed in the same order as in Table 7. The selected model is associated with the lowest MSE and the MSE is mostly in the same ranking as the model plausibility. This corroborates the generalization capability of the selected model and the effectiveness for model selection by the proposed method.

6. CONCLUSION

In this paper we proposed a novel Bayesian nonparametric general regression method. The general regression was reformulated under the Bayesian framework and this proposed method does not require specification of an explicit parametric model for the underlying input-output relationship. Moreover, there are only two unknown parameters regardless of the number of design variables, namely, the smoothing scale parameter and the prediction-error scale parameter. Furthermore, the method is able to identify the design variables necessary for the modeling of the quantity of interest. The illustrative examples demonstrated the success of the proposed methodology for the identification of significant design variables and for reliable prediction. The proposed approach exhibits two major appealing advantages in comparison to other regression techniques. First, it resolves the subjectivity problem in Bayesian model selection for traditional regression due to the prior distribution of the regression parameters. Second, the number of model candidates is drastically reduced so it allows for the examination of a much larger number of potential design variables.

ACKNOWLEDGMENT

This investigation has been supported by the Research Committee of the University of Macau under Research Grant MYRG2015-00048-FST. This generous support is gratefully acknowledged.

REFERENCES

1. Kasparis, I., Andreou, E., and Phillips, P. C. B., Nonparametric predictive regression, *J. Econometrics*, vol. **185**, no. 2, pp. 468–494, 2015.
2. Donnelly, A., Misstear, B., and Broderick, B., Real time air quality forecasting using integrated parametric and non-parametric regression techniques, *Atmos. Environ.*, vol. **103**, pp. 53–65, 2015.

3. Takeda, H., Farsiu, S., and Milanfar, P., Kernel regression for image processing and reconstruction, *IEEE Trans. Image Process.*, vol. **16**, no. 2, pp. 349–366, 2007.
4. Clark, S., Traffic prediction using multivariate nonparametric regression, *J. Transp. Eng.*, vol. **129**, no. 2, pp. 161–168, 2003.
5. Yanfeng, G., Chen, W., Baoxue, L., and Ye, Z., A kernel-based nonparametric regression method for clutter removal in infrared small-target detection applications, *IEEE Geosci. Remote Sens. Lett.*, vol. **7**, no. 3, pp. 469–473, 2010.
6. Lopes, H., Barcellos, J., Kubrusly, J., and Fernandes, C., A non-parametric method for incurred but not reported claim reserve estimation, *Int. J. Uncertainty Quantif.*, vol. **2**, no. 1, pp. 39–51, 2012.
7. Specht, D. F., A general regression neural network, *IEEE Trans. Neural Networks*, vol. **2**, no. 6, pp. 568–576, 1991.
8. Nadaraya, E. A., On estimating regression, *Theory Probab. Appl.*, vol. **9**, no. 1, pp. 141–142, 1964.
9. Watson, G. S., Smooth regression analysis, *Sankhya, Ser. A*, vol. **26**, no. 4, pp. 359–372, 1964.
10. Leung, M. T., Chen, A.-S., and Daouk, H., Forecasting exchange rates using general regression neural networks, *Comput. Oper. Res.*, vol. **27**, no. 11, pp. 1093–1110, 2000.
11. Ben-Nakhi, A. E. and Mahmoud, M. A., Cooling load prediction for buildings using general regression neural networks, *Energy Convers. Manage.*, vol. **45**, no. 13, pp. 2127–2141, 2004.
12. Li, C., Bovik, A. C., and Wu, X., Blind image quality assessment using a general regression neural network, *IEEE Trans. Neural Networks*, vol. **22**, no. 5, pp. 793–799, 2011.
13. Wehenkel, L., Contingency severity assessment for voltage security using non-parametric regression techniques, *IEEE Trans. Power Syst.*, vol. **11**, no. 1, pp. 101–111, 1996.
14. Kiefa, M. A., General regression neural network for driven piles in cohesionless soils, *J. Geotech. Geoenviron. Eng.*, vol. **124**, no. 12, pp. 1177–1185, 1998.
15. Ching, J., Beck, J. L., Porter, K. A., Bayesian state and parameter estimation of uncertain dynamical systems, *Probab. Eng. Mech.*, vol. **21**, no. 1, pp. 81–96, 2006.
16. Yuen, K. V. and Katafygiotis, L. S., Model updating using noisy response measurements without knowledge of the input spectrum, *Earthquake Eng. Struct. Dyn.*, vol. **34**, no. 2, pp. 167–187, 2005.
17. Mu, H. Q. and Yuen, K. V., Novel outlier-resistant extended Kalman filter for robust online structural identification, *J. Eng. Mech.*, vol. **141**, no. 1, 04014100, 2015.
18. Yuen, K. V., Liang, P. F., and Kuok, S. C., Online estimation of noise parameters for Kalman filter, *Struct. Eng. Mech.*, vol. **47**, no. 3, pp. 361–381, 2013.
19. Yuen, K. V. and Kuok, S. C., Efficient Bayesian sensor placement algorithm for structural identification: a general approach for multi-type sensory systems, *Earthquake Eng. Struct. Dyn.*, vol. **44**, no. 5, pp. 757–774, 2015.
20. Ortiz, G. A., Alvarez, D. A., and Bedoya-Ruiz, D., Identification of Bouc-Wen type models using the transitional Markov chain Monte Carlo method, *Comput. Struct.*, vol. **146**, no. 1, pp. 252–269, 2015.
21. Huang, Y. and Beck, J. L., Hierarchical sparse Bayesian learning for structural health monitoring with incomplete model data, *Int. J. Uncertainty Quantif.*, vol. **5**, no. 2, pp. 139–169, 2015.
22. Lei, Y., Li, Q., Chen, F., and Chen, Z., Damage identification of frame structures with joint damage under earthquake excitation, *Adv. Struct. Eng.*, vol. **17**, no. 8, pp. 1075–1088, 2014.
23. Lei, Y., Zhou, H., and Liu, L. J., An on-line integration technique for structural damage detection and active optimal vibration control, *Int. J. Struct. Stab. Dyn.*, vol. **14**, no. 5, 1440003, 2014.
24. Ching, J., Chen, J. R., Yeh, J. Y., and Phoon, K. K., Updating uncertainties in friction angles of clean sands, *J. Geotech. Geoenviron. Eng.*, vol. **138**, no. 2, pp. 217–229, 2012.
25. Yan, W. M., Yuen, K. V., and Yoon, G. L., Bayesian probabilistic approach for the correlations of compressibility index for marine clays, *J. Geotech. Geoenviron. Eng.*, vol. **135**, no. 12, pp. 1932–1940, 2009.
26. Chiu, C. F., Yan, W. M., and Yuen, K. V., Reliability analysis of soil-water characteristics curve and its application to slope stability analysis, *Eng. Geol.*, vol. **135**, pp. 83–91, 2012.
27. Hoi, K. I., Yuen, K. V., and Mok, K. M., Prediction of daily average PM_{10} concentrations by statistical time-varying model, *Atmos. Environ.*, vol. **43**, no. 16, pp. 2579–2581, 2009.
28. Field, R. V. Jr., Constantine, P., and Boslough, M., Statistical surrogate models for prediction of high-consequence climate change, *Int. J. Uncertainty Quantification*, vol. **3**, no. 4, pp. 341–355, 2013.

29. Beck, J. L. and Taflanidis, A. A., Prior and posterior robust stochastic predictions for dynamical systems using probability logic, *Int. J. Uncertainty Quantification*, vol. **3**, no. 4, pp. 271–288, 2013.
30. Poulakis, Z., Valougeorgis, D., and Papadimitriou, C., Leakage detection in water pipe networks using a Bayesian probabilistic framework. *Probab. Eng. Mech.*, vol. **18**, no. 4, pp. 315–327, 2003.
31. Ginting, V., Pereira, F., and Rahunanthan, A., A multi-stage Bayesian prediction framework for subsurface flows, *Int. J. Uncertainty Quantification*, vol. **3**, no. 6, pp. 499–522, 2013.
32. Papadimitriou, C., Beck, J. L., and Katafygiotis, L. S., Updating robust reliability using structural test data, *Probab. Eng. Mech.*, vol. **16**, no. 2, pp. 103–113, 2001.
33. Ching, J. and Hsieh, Y. H., Local estimation of failure probability function and its confidence intervals with maximum entropy principle, Probab. Eng. Mech., vol. **22**, no. 1, pp. 39–49, 2007.
34. Der Kiureghian, A., Analysis of structural reliability under parameter uncertainties, *Probab. Eng. Mech.*, vol. **23**, no. 4, pp. 351–358, 2008.
35. Jensen, H. A., Vergara, C., Papadimitriou, C., and Millas, E., The use of updated robust reliability measures in stochastic dynamical systems, *Comput. Methods Appl. Mech. Eng.*, vol. **267**, no. 1, pp. 293–317, 2013.
36. Parzen, E., On estimation of a probability density function and mode. *Ann. Math. Stat.*, vol. **33**, no. 3, pp. 1065–1076, 1962.
37. Cacoullos, T., Estimation of a multivariate density, *Ann. Inst. Stat. Math.*, vol. **18**, no. 1, pp. 179–189, 1966.
38. Beck, J. L., Bayesian system identification based on probability logic, *Struct. Control Health Monit.*, vol. **17**, no. 7, pp. 825–847, 2010.
39. Yuen, K. V., *Bayesian Methods for Structural Dynamics and Civil Engineering*, New York: John Wiley & Sons, Inc., 2010.
40. Yuen, K. V., Hoi, K. I., and Mok, K. M., Selection of noise parameters for Kalman filter, *Earthquake Eng. Eng. Vibr.*, vol. **6**, no. 1, pp. 49–56, 2007.
41. Papadimitriou, C., Beck, J. L., and Katafygiotis, L. S., Asymptotic expansions for reliability and moments of uncertain systems, *J. Eng. Mech.*, vol. **123**, no. 12, pp. 1219–1229, 1997.
42. Yuen, K. V. and Mu, H. Q., Real-time system identification: an algorithm for simultaneous model class selection and parametric identification, *Comput.-Aided Civil Infrastruct. Eng.*, vol. **30**, no. 10, pp. 785–801, 2015.
43. Smola, A. J. and Schölkopf, B., Bayesian kernel methods, In Mendelson, S., Smola, A. J., Eds., *Advanced Lectures on Machine Learning*, Lecture Notes in Computer Science, vol. **2600**, pp. 65–117, 2003.
44. AI FAQ/Neural Nets index, Part 2 of 7: Learning. Available at http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html, 2015.
45. Sanli, S. G., Kizilkanat, E. D., Boyan, N., Ozsahin, E. T., Bozkir, M. G., Soames, R., Erol, H., and Oguz, O., Stature estimation based on hand length and foot length, *Clin. Anat.*, vol. **18**, no. 8, pp. 589–596, 2005.
46. Winner, L., Department of Statistics, University of Florida: Miscellaneous datasets. Available at http://www.stat.ufl.edu/~winner/datasets.html, 2015.
47. Ashizawa, K., Kumakura, C., Kusumoto, A., and Narasaki, S., Relative foot size and shape to general body size in Javanese, Filipinas and Japanese with special reference to habitual footwear types, *Ann. Human Biol.*, vol. **24**, no. 2, pp. 117–129, 1997.
48. Gordon, C. C. and Buikstra, J. E., Linear models for the prediction of stature from foot and boot dimensions, *J. Forensic Sci.*, vol. **37**, no. 3, pp. 771–782, 1992.