A weighted heteroscedastic Gaussian Process Modelling via particle swarm optimization

Xiaodan Hong [a], Yongsheng Ding [a,**], Lihong Ren [a], Lei Chen [a], Biao Huang [b,*]

[a] Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, College of Information Sciences and Technology, Donghua University, Shanghai 201620, PR China
[b] Department of Chemical and Materials Engineering, University of Alberta, Canada

ARTICLE INFO

ABSTRACT

In many chemical engineering applications, it is often difficult to get accurate first-principle models because of complexity of modern processes. Even if it is possible to do so, it is often time consuming and computationally expensive. Hence, there is a growing need to develop data-driven models. Gaussian process regression (GPR) model has been extensively applied in data-based modelling due to its good adaptability to deal with high dimensional, small samples, and nonlinear problems. The standard GPR algorithm assumes constant noise power throughout the sampling process. However, in process systems, the observation noise often varies so that different sample points are corrupted by different degrees of noise. Under these circumstances, the standard GPR algorithm may not work properly. To model Gaussian process with heteroscedastic noise, this paper introduces a weighting strategy into the standard GPR algorithm, and proposes three weighted GPR algorithms: the clustered GPR (C-GPR) algorithm, the partial weighted GPR (PW-GPR) algorithm and the weighted GPR (W-GPR) algorithm. Different from the standard GPR algorithm, three weighted algorithms put the weight on sampled data by calculating the noise variance for each data point. In addition, in order to optimize the proposed algorithms, this paper utilizes the particle swarm optimization (PSO) algorithm to estimate hyper-parameters of the GPR model, instead of using the traditional conjugate gradient (CG) method. The effectiveness of the three weighted GPR algorithms is verified by means of two numerical examples and a wet spinning coagulation process. Extensive simulation results demonstrate that the proposed algorithms optimized by the PSO algorithm can improve prediction accuracy of the GPR model.

1. Introduction

In many industrial sectors, such as the chemical, automotive and semiconductor industries, it is common to rely on mathematical/computer models to guide process operations [1–4]. However, it is often difficult to get accurate first-principle models because of complexity of modern industries. Even if the complex mechanism of the system is known, it is often time consuming and expensive to build a detailed model [5]. For such systems, since it is difficult to derive the first-principle models, data-driven modelling approaches, such as neural networks, support vector regression and Gaussian process regression (GPR) modelling, can provide a good alternative [6–10].

The Gaussian process regression (GPR) algorithm is a recently developed nonlinear modelling method in the field of machine learning. It has sound theoretical basis in Statistics, and it has a good adaptability to high dimension, small sample set, and nonlinear problems, as well as strong generalization ability [11]. In addition, compared to other regression models such as support vector machine, fuzzy model, and neural networks, GPR model is easy to implement, its hyper-parameters could be adaptively acquired, and output prediction is associated with probability distribution [10–12]. Due to these advantages, GPR algorithms have gained rapid development and attracted a lot of research interests. In recent years, GPR has become one of the main focuses of research in the field of machine learning [12–14], and it has been successfully used in many applications [15–17].

In many real industrial processes, it is impossible to get exact values of process variables. The collected data are usually accompanied by noise. The standard GPR algorithm assumes constant noise power throughout

---

* Corresponding author.
** Corresponding author.
   E-mail addresses: 1773814891@qq.com (X. Hong), ysding@dhu.edu.cn (Y. Ding), lhren@dhu.edu.cn (L. Ren), leichen@dhu.edu.cn (L. Chen), bhuang@ualberta.ca (B. Huang).

the whole sampling process. It assumes each output measurement is affected by homoscedastic Gaussian white noise, which has a constant variance everywhere [18,19]. This assumption of homoscedastic Gaussian noise makes GPR inference to be more computationally tractable. But, it has limitation when applied to many practical engineering problems. In process systems, owing to some disturbances, such as the malfunction of sensors and incorrect measuring by technicians, output measurements may corrupted with heteroscedastic Gaussian white noise [20,21]. In this heteroscedastic noise scenario, the effectiveness of the prediction can be weakened if we assume variance of each observation to be the same while the reality is otherwise. To deal with the heteroscedastic noise problem mentioned above, this paper develops three different weighting strategies, on the basis of the standard GPR method, and proposes the clustered GPR (C-GPR) algorithm, the partial weighted GPR (PW-GPR) algorithm and the weighted GPR (W-GPR) algorithm. The proposed weighted GPRs improve prediction effectiveness and prediction accuracy of the GPR model since they consider the GPR model with the varying noise variances which evaluated by employing repeated sampling, and it lead to that different sample has different weight in the GPR modelling. This paper is the significantly extended version of our conference paper [22] which focuses only on the W-GPR algorithm.

Optimization of hyper-parameters is an important part of modelling Gaussian process, and it is related to the reliability of the regression model. The conjugate gradient (CG) algorithm is one of the most commonly-used methods [11,13,23] to estimate parameters. But, this optimization method relies heavily on the selection of initial guess, and it is difficult to determine the number of iterations. In addition, for most of the GPR estimation, optimization of hyper-parameters is not a convex optimization problem. Thus, the gradient based method can easily result in local optimum [24,25]. To increase prediction accuracy of the proposed GPR models, this paper considers a swarm intelligence algorithm - particle swarm optimization (PSO) algorithm [26,27], instead of the traditional gradient method, for the optimization of hyper-parameters of the GPR model. The PSO algorithm is a kind of evolutionary algorithm based on the flock foraging swarm intelligence relationship [26]. Compared with the CG algorithm, this method does not rely on the selection of the initial guess and is simple and feasible in addressing the problem of global optimization since it has strong global search ability and its information sharing mechanism is helpful for a rapid convergence [28–30]. Also, it has been proved to improve the prediction accuracy of the standard GPR model [28]. Due to these advantages, the PSO algorithm will be used to estimate the optimal parameters of the proposed GPR models by finding the maximum likelihood.

The main contributions of this paper are as follows: (i) three different weighting strategies are developed in the GPR model for different noise scenarios, respectively, and the C-GPR, the PW-GPR and the W-GPR algorithms are proposed with their application scope; (ii) the PSO algorithm is utilized to optimize the proposed weighted GPR models.

The rest of the paper is outlined as follows. Section 2 introduces the principle of the standard GPR. Section 3 proposes three GPR algorithms with weighting strategy, namely the C-GPR algorithm, the PW-GPR algorithm, and the W-GPR algorithm. Section 4 uses the PSO algorithm to search the optimal hyper-parameters of the GPR model. Simulation examples are given in Section 5. The conclusion is provided in Section 6.

2. The standard GPR algorithm

GPR algorithm is among the functional approximation methods of supervised learning in the field of machine learning. The algorithm is used to evaluate the distribution of value function with the sample data. From the standpoint of function space, we describe the distribution of a function by defining a Gaussian process (GP). Further, we can directly conduct Bayesian inference in this function space [11,12,31]. The formal definition of the GP is given below.

Definition 1. [12]. A GP is a collection of random variables, any finite number of which has (consistent) joint Gaussian distributions.

Indeed, a GP is completely determined by its mean function and covariance function, that is to say, if mean function $m(X)$ and covariance function $k(X, X\prime)$ are known, we can fully determine the GP. Obviously, if the mean and covariance are a vector and matrix respectively, we will have the following functional distribution:

$$GP : f(X) \sim GP\big(m(X), k(X, X^{'})\big) \tag{1}$$

where, $m(X) = E[f(X)]$, $k(X, X') = E[(f(X) - m(X))(f(X') - m(X))]$. $f(X)$ represents the underlying function value at the input X, and the n function values $F = \{f(X_1), \cdots, f(X_n)\}$ can be assumed as a n-variates joint Gaussian distribution. For convenience, we usually assume that mean of an arbitrary point of this joint Gaussian distribution is 0. This can be achieved by mean centering data in practice. Thus, the correlation between different sample points only depends on the covariance function $k(X, X')$. In GPR, we often choose 'square exponential function' to describe the covariance function as follows:

$$k(X, X') = \sigma_f^2 \exp\left(-\frac{(X - X')^2}{2l^2}\right) \tag{2}$$

where, The length parameter l determines how the distance between variables X and X' affects the correlation between them. σ_f^2 is a variance term of the covariance function. If $X = X'$, then $k(X, X')$ approaches the maximum correlation, which means $f(X)$ is nearly perfectly correlated with $f(X')$. If X is far away from X', we have instead $k(X, X') \approx 0$, meaning almost no correlation between two points.

Consider the standard GPR model with noise:

$$y(X) = f(X) + \varepsilon \tag{3}$$

where, f is an underlying function, and function value $f(X)$ conforms to the distribution described by Equation (1). X is the input variable, and $y(X)$ is the observed value corrupted by noise. In standard GPR, ε is assumed as the homoscedastic Gaussian white noise which has the following distribution:

$$\varepsilon \sim N\big(0, \sigma_n^2\big) \tag{4}$$

The main objective of GPR modelling is to evaluate the function value $f(X_*)$ at the new input X_*. In fact, for multivariate input variables, the covariance function in Equation (2) can be expressed in a more general form as:

$$k(X, X') = \sigma_f^2 \exp\left(-\sum_{h=1}^{d} \frac{(x_h - x'_h)^2}{2l_h^2}\right) \tag{5}$$

where, d is the dimension of the input X, x_h and x'_h, $h = 1, \cdots, d$ are h-th dimension of inputs $X \in \mathbb{R}^d$ and $X' \in \mathbb{R}^d$, respectively. l_h represent the characteristic length-scale corresponding to the h-th input variable dimension.

Consider a training sample set $S = \{(X_i, y_i), i = 1, \cdots, n\}$, where n is the number of the given sample, $X_i \in \mathbb{R}^d$ is the input variable vector at sampling instant i, and $y_i \in \mathbb{R}$ is the corresponding output variable. Now, given a new input X_*, we need to estimate the underlying function value $f_* = f(X_*)$ by a relationship established in the data set S. According to assumptions for Gaussian process described above, the function value $F = \{f(X_1), \cdots, f(X_n)\}$ follows joint Gaussian distributions, that is, $F \sim N(0, Var(F))$. Hence, n observations $Y = \{y_1, \cdots, y_n\}$ together with the new function value f_* can be considered as a multivariate Gaussian process. The distribution of Y can be written as below:

$$Y \sim N\big(0, Var(F) + \sigma_n^2 I_n\big) \tag{6}$$

and the joint distribution between Y and f_* is: