European Journal of Operational Research 228 (2013) 131–140

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Error estimation properties of Gaussian process models in stochastic simulations

Andres F. Hernandez, Martha A. Grover *

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0100, United States

ARTICLE INFO

Article history:
Received 5 June 2012
Accepted 25 December 2012
Available online 29 January 2013

Keywords:
Simulation
Regression
Metamodeling
Gaussian process models
Error estimation

ABSTRACT

The theoretical relationship between the prediction variance of a Gaussian process model (GPM) and its mean square prediction error is well known. This relationship has been studied for the case when deterministic simulations are used in GPM, with application to design of computer experiments and metamodeling optimization. This article analyzes the error estimation of Gaussian process models when the simulated data observations contain measurement noise. In particular, this work focuses on the correlation between the GPM prediction variance and the distribution of prediction errors over multiple experimental designs, as a function of location in the input space. The results show that the error estimation properties of a Gaussian process model using stochastic simulations are preserved when the signal-to-noise ratio in the data is larger than 10, regardless of the number of training points used in the metamodel. Also, this article concludes that the distribution of prediction errors approaches a normal distribution with a variance equal to the GPM prediction variance, even in the presence of significant bias in the GPM predictions.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

A recurring question in metamodeling is how to assess the prediction accuracy of the model. Even though a validation step is typically performed as part of the metamodel construction, the question remains as to how accurate is the prediction of the model at an untried sample point. Assume that the response of an expensive simulation f can be described as $y_{tr}(\mathbf{x}) = \mathrm{f}(\mathbf{x})$, where the subscript *tr* denotes the true mean response of the simulation at $\mathbf{x}$. The approximate model of the expensive simulation, $\hat{\mathrm{f}}$, predicts the mean response at $\mathbf{x}$ as: $y_{app}(\mathbf{x}) = \hat{\mathrm{f}}(\mathbf{x})$. Therefore, the mean square prediction error of the approximate model at $\mathbf{x}$ is:

$$\delta^2(\mathbf{x}) = [y_{tr}(\mathbf{x}) - y_{app}(\mathbf{x})]^2 \tag{1}$$

In the scenario of expensive simulations, a user may have limited resources (i.e. limited number of simulations) for the training and validation steps of the metamodel. This situation is more difficult when the user is approximating *stochastic* simulations, since each observed response y from the simulation is corrupted by a measurement noise η around the mean response $y(\mathbf{x}) = y_{tr}(\mathbf{x}) + \eta$. Because of this scenario, researchers have looked for alternatives to obtain an estimation of the prediction error that do not require additional evaluations from the expensive simulation.

Gaussian process modeling (GPM) is one of the most popular methods for constructing approximate models, not only because of its flexibility and good prediction results, but also because it has its own error estimation on the GPM prediction. According to the theory of GPM, when the GPM structure is completely known (that is, the true GPM parameter set, the true local correlation structure and the true regression functions in the model are known), the GPM prediction variance is an error estimator of the mean square prediction error. In practice, the user has to make decisions about the GPM structure, incurring an additional model uncertainty factor that alters this theoretical property. Many of the applications of GPM implement a "plug-in" version of the Gaussian process, using an estimated parameter set. Santner et al. (2003) called this version of GPM the *empirical* best linear unbiased predictor. Several authors already mentioned this situation, suggesting that the GPM prediction variance of the empirical GPM version is an underestimator of the "true" GPM prediction variance (Cressie, 1993; den Hertog et al., 2006). These issues raise questions about how to estimate the error in Gaussian process models in practice.

Error estimation measures are useful for the assessment of approximate models. In many applications, the success of approximate models depends on the accuracy in the error estimation. Error estimators are used to quantify the level of uncertainty or "trust" in the prediction of an approximate model, therefore indicating particular regions in the input space where additional samples are required. Some examples include improvement of design of experiments in the creation of approximate models (Hernandez and Grover, 2010), and optimization of time-consuming computer simulations via black-box models (Kleijnen et al., 2010).

* Corresponding author. Tel.: +1 404 894 2878; fax: +1 404 894 2866.
E-mail address: martha.grover@chbe.gatech.edu (M.A. Grover).
URL: http://grover.chbe.gatech.edu/ (M.A. Grover).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.12.033

Early findings about error estimation of GPM were made by Meckesheimer et al. (2002). They evaluate leave-k-out cross-validation strategies as a procedure to assess the accuracy of low-order polynomial functions, radial basis functions and kriging models over the design space. Goel et al. (2009) presented a detailed study on error estimation, evaluating response surfaces and kriging models for six classical benchmark examples in the statistics field. As a result of this last study, Goel et al. concluded that local evaluations of GPM prediction variance can be used for global error estimation of Gaussian process models. As relevant as these results are in the error estimation of GPM, the studies were limited to predictions of deterministic simulations.

The presence of noise in the observations incorporates an additional element in the GPM prediction that has been discussed previously in the literature. Kleijnen and coworkers have worked extensively in the use of kriging models for random simulations (Kleijnen et al., 2010; Kleijnen and van Beers, 2005; van Beers and Kleijnen, 2003, 2008), using replicates at each sample point to calculate sample means, and then treating those values as deterministic outcomes in the GPM construction. In the area of error estimation, the same group implemented a parametric bootstrapping approach to calculate the mean square prediction error of the GPM (den Hertog et al., 2006), but it was not used as an error estimator in the approximate model and it was only employed with deterministic simulations. A similar implementation of this bootstrapping approach was also used to evaluate the uncertainty of time-course experimental data in cell signaling pathways and network topology of time-series gene expression data (Kirk and Stumpt, 2009). Ankenman et al. (2010) extended Kleijnen's GPM for random simulations with their stochastic kriging model, which models the intrinsic uncertainty, or noise, in the simulations with an additional variance parameter for each sample point. Different from these papers, where the major interest was the GPM mean prediction, the work presented here focuses on the GPM prediction variance and its role as an error estimator of the approximate model when stochastic observations are used in GPM.

2. Theory

Consider a set $\mathcal{D}$ of n input/output pairs $\{\mathbf{x}_i, y(\mathbf{x}_i)\}$, where $\mathbf{x}_i \in \mathbb{R}^d,\ y(\mathbf{x}_i) \in \mathbb{R},\ i = 1, \ldots, n$. This set of input/output pairs will be referred to as sample points or experimental points. Consider also that the observed data $y(\mathbf{x}_i)$ contains an additive measurement noise $\eta \sim \mathcal{N}(0, \sigma_u^2)$ along with the true response of the simulation $y_{tr}(\mathbf{x}_i) \in \mathbb{R}$. The subscript u represents the uncorrelated nature of the additive noise in each of the experimental points.

$$y(\mathbf{x}_i) = y_{tr}(\mathbf{x}_i) + \eta \tag{2}$$

Despite the presence of noise in the observed data, the objective is to predict the true response of the simulation $y_{tr}(\mathbf{x})$ at some other point $\mathbf{x}$.

2.1. Gaussian process model (GPM)

In a GPM, the set of unknown true responses of the simulations are treated as random variables that are drawn from a joint Gaussian distribution (Rasmussen and Williams, 2006). For the elements in the set $\mathcal{D}$, the mean function $m(\mathbf{x})$ and the covariance matrix $K \in \mathbb{R}^{n\times n}, K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ of the true response values $y_{tr}(\mathbf{x}_i)$ are defined as

$$\mathbf{y}_{tr} \sim \mathcal{GP}(m, K) \tag{3}$$

$$\mathbb{E}[\mathbf{y}_{tr}] = m(\mathbf{x}) = H\boldsymbol{\beta} \tag{4}$$

$$\mathbb{E}[(\mathbf{y}_{tr} - H\boldsymbol{\beta})(\mathbf{y}_{tr} - H\boldsymbol{\beta})^T] = K \tag{5}$$

where $\mathbf{y}_{tr} = [y_{tr}(\mathbf{x}_1)\ y_{tr}(\mathbf{x}_2) \ldots y_{tr}(\mathbf{x}_n)]^T, \mathbf{y}_{tr} \in \mathbb{R}^n,\ H \in \mathbb{R}^{n\times p}$ represents a set of p regression or trend functions evaluated at the $\mathbf{x}_i$ inputs in $\mathcal{D}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the p regression coefficients.

Based on the description of the observed data in Eq. (2), and the multivariate distribution of the true response, the output/observed information $y(\mathbf{x}_i)$ in the set $\mathcal{D}$ can alternatively be drawn from a multivariate Gaussian distribution as

$$\mathbf{y} \sim \mathcal{GP}(m, K + \sigma_u^2 I) \tag{6}$$

$$\mathbb{E}[\mathbf{y}] = m(\mathbf{x}) = H\boldsymbol{\beta} \tag{7}$$

$$\mathbb{E}[(\mathbf{y} - H\boldsymbol{\beta})(\mathbf{y} - H\boldsymbol{\beta})^T] = K + \sigma_u^2 I \tag{8}$$

where $\mathbf{y} = [y(\mathbf{x}_1)\ y(\mathbf{x}_2) \ldots y(\mathbf{x}_n)]^T, \mathbf{y} \in \mathbb{R}^n$ and $I \in \mathbb{R}^{n\times n}$ is the identity matrix.

The covariance matrix K can be constructed only after defining the function k that models the correlation between the true values of the residual. Basically, the correlation is described using the distance between sample points, with a monotonically decaying function. A common distance-based correlation function employed in the GPM literature is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_c^2 \exp\left[-\frac{1}{2}\sum_{a=1}^{d}\frac{(x_{i,a} - x_{j,a})^2}{\ell_a^2}\right] \tag{9}$$

where $\boldsymbol{\theta} = (\ell_1^2 \ldots \ell_d^2, \sigma_c^2, \sigma_u^2)$ are the parameters that control the features of the correlation between samples in the GPM. The subscript c is used to describe a correlated variance in the model that is weighted using the distance-based correlation function. Based on this correlation function, the GPM can also be understood as an empirical model that uses a positive-definite function to model the correlation between the residuals of a linear-in-parameters model and the true responses (Cressie, 1993; Koehler and Owen, 1996).

The mathematical description of GPM as a regression model can be obtained by solving a constrained optimization problem for the best linear unbiased predictor and its mean square error (Goldberger, 1962). This constrained optimization problem is solved using the method of Lagrange multipliers, and its solution is known as the *best linear unbiased predictor (BLUP)*, which corresponds to:

$$\hat{y}(\mathbf{x}, \mathcal{D}) = \mathbf{h}^T(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{k}^T(\mathbf{x}, \mathcal{D})\left(K + \sigma_u^2 I\right)^{-1}[\mathbf{y} - H\hat{\boldsymbol{\beta}}] \tag{10}$$

where $\mathbf{k}(\mathbf{x}, \mathcal{D}) \in \mathbb{R}^n$ is the correlation vector between $\mathbf{x}$ and each of the $\mathbf{x}_i$ samples in $\mathcal{D}$ using a correlation function such as the one in Eq. (9); $\mathbf{h} \in \mathbb{R}^p$ represents a set of p regression functions evaluated at $\mathbf{x}$ and

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(H^T\left(K + \sigma_u^2 I\right)^{-1} H\right)^{-1} H^T\left(K + \sigma_u^2\ I\right)^{-1}\mathbf{y} \tag{11}$$

is the generalized least-squares estimator of the regression coefficients. Similarly, the GPM prediction variance of the linear predictor is calculated as

$$\sigma_y^2(\mathbf{x}, \mathcal{D}) = k(\mathbf{x}, \mathbf{x}) - [\mathbf{h}^T(\mathbf{x})\quad \mathbf{k}^T(\mathbf{x}, \mathcal{D})]\begin{bmatrix} 0 & H^T \\ H & K + \sigma_u^2\ I \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{h}(\mathbf{x}) \\ \mathbf{k}(\mathbf{x}, \mathcal{D}) \end{bmatrix} \tag{12}$$

where $k(\mathbf{x}, \mathbf{x})$ is the evaluation of the correlation function between the unknown point and itself. It is also important to notice that in the original description of this constrained optimization (Goldberger, 1962), Goldberger did not specify a correlation function to describe the regression covariance matrix. Goldberger does not describe the nature of the observations to be either deterministic or stochastic. This means that either $\mathbf{y}$ or $\mathbf{y}_{tr}$ can be used in the GPM equations. When the deterministic simulations $\mathbf{y}_{tr}$ are used in GPM, the parameter σ_u^2 has a value of 0, and Eqs. (10)–(12) are modified accordingly.

Download English Version:

https://daneshyari.com/en/article/6898059

Download Persian Version:

https://daneshyari.com/article/6898059

Daneshyari.com