Computational Statistics and Data Analysis 73 (2014) 40–52

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Efficient optimization of the likelihood function in Gaussian process modelling☆

A. Butler [a], R.D. Haynes [a,*], T.D. Humphries [a], P. Ranjan [b]

[a] Department of Mathematics and Statistics, Memorial University, St. John's, NL, Canada, A1C 5S7

[b] Department of Mathematics and Statistics, Acadia University, Wolfville, NS, Canada, B4P 2R6

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form 22 November 2013
Accepted 24 November 2013
Available online 4 December 2013

Keywords:
BFGS
Clustering
Computer simulators
Dividing rectangles
Implicit filtering
Ill-conditioning
Nugget

ABSTRACT

Gaussian Process (GP) models are popular statistical surrogates used for emulating computationally expensive computer simulators. The quality of a GP model fit can be assessed by a goodness of fit measure based on optimized likelihood. Finding the global maximum of the likelihood function for a GP model is typically challenging, as the likelihood surface often has multiple local optima, and an explicit expression for the gradient of the likelihood function may not be available. Previous methods for optimizing the likelihood function have proven to be robust and accurate, though relatively inefficient. Several likelihood optimization techniques are proposed, including two modified multi-start local search techniques, that are equally as reliable, and significantly more efficient than existing methods. A hybridization of the global search algorithm Dividing Rectangles (DIRECT) with the local optimization algorithm BFGS provides a comparable GP model quality for a fraction of the computational cost, and is the preferred optimization technique when computational resources are limited. Several test functions and an application motivated by oil reservoir development are used to test and compare the performance of the proposed methods with the implementation provided in the R library **GPfit**. The proposed method is implemented in a `Matlab` package, **GPMfit**.

1. Introduction

Computer simulators are useful tools for modelling complex real world systems that are either impractical, expensive, or time consuming to physically observe. For example, the energy generated by the tides of large ocean basins (Greenberg, 1979), the estimation of the magnetic field generated near the Milky Way (Short et al., 2007), and the analysis of the flow of oil in a reservoir (Aziz and Settari, 1979) — the latter of which motivated this research — can be achieved through the use of computer simulators. That being said, realistic computer simulators can be computationally expensive to run, and as a result are often emulated using statistical models, such as Gaussian Process (GP) models (Sacks et al., 1989).

The maximum likelihood approach for fitting a GP model to deterministic simulator output requires minimizing the negative log-likelihood, or deviance. Rasmussen and Williams (2006) proposed the use of either a randomized multi-start conjugate gradient method or Newton's method for this problem. Explicit information about the gradient of deviance cannot be easily obtained, however, and the deviance function surface often has many local optima, making the optimization

☆ Supplementary Material: The open source `Matlab` package **GPMfit** is available for download on `SourceForge.net`. See `Readme.txt` for detailed instruction. The main functions are `model_fit.m` and `predictor_iterative.m`.

* Corresponding author. Tel.: +1 7098648825.
E-mail addresses: adsb85@mun.ca (A. Butler), rhaynes@mun.ca, rhaynes74@gmail.com (R.D. Haynes), thumphries@mun.ca (T.D. Humphries), pritam.ranjan@acadiau.ca (P. Ranjan).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.11.017

problem challenging (MacDonald et al., 2013). Derivative-free optimization techniques, such as the genetic algorithm used by Ranjan et al. (2011), or the differential evolution algorithm used by Petelin et al. (2011), are robust, but can be computationally inefficient. Gradient approximation methods, such as the Broyden–Fletcher–Goldfarb–Shanno method (BFGS) (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), are generally faster, but have the potential to converge only locally if poorly initialized. MacDonald et al. (2013) proposed a clustering-based multi-start BFGS algorithm, which allows for a more global search to be performed. Nonetheless, this method requires multiple executions of BFGS, which is also computationally expensive.

In this paper we investigate several optimization techniques in order to improve the efficiency of the likelihood optimization process. Each technique is a combination of global and local search strategies. At the global level, we propose using the Dividing Rectangles algorithm (DIRECT) (Finkel, 2003) as an alternative to the clustering-based approach for choosing the starting point(s) of the local search. In terms of the local search, we compare the performance of BFGS with that of Implicit Filtering (IF), a sophisticated pattern search algorithm developed by Kelley (2011) for multimodal noisy functions. We use several test functions and an application motivated by real-world oil reservoir development to compare the performance of different optimization techniques, as measured by the prediction accuracy (optimized deviance and root mean squared prediction error) and number of deviance function evaluations (FEs) required to optimize the deviance. After an extensive case study we find that a hybrid approach of DIRECT and BFGS is the most efficient optimization technique for fitting such GP models.

The remainder of the paper is outlined as follows. Section 2 describes the GP model and the main components of the newly developed `Matlab` package **GPMfit**. In Section 3 we briefly outline the optimization techniques used for minimizing the deviance. Section 4 provides the results and analysis for several test functions, followed by an example in Section 5 where the GP model is fit to an oil reservoir simulator using our proposed method. Concluding remarks are provided in Section 6.

2. The Gaussian process model

The GP model requires as input a set of design points, $x_i = (x_{i1}, \ldots, x_{id})'$, and the corresponding simulator outputs, $y_i = y(x_i)$, where $i = 1, \ldots, n$, and n is the number of user supplied design points. Here, the prime symbol, $'$, denotes the transpose of vectors or matrices. We assume that the simulator provides a scalar valued output, y_i, for each d-dimensional design point x_i, and we use $Y = (y_1, \ldots, y_n)'$ to denote the $n \times 1$ vector of simulator outputs. The simulator output is modelled as

$$y_i = \mu + z(x_i),$$

where μ is the overall mean, and $z(x_i)$ is a GP with $E[z(x_i)] = 0$, $\mathrm{Var}[z(x_i)] = \sigma^2$, and $\mathrm{Cov}[(z(x_i), z(x_j))] = \sigma^2 R_{ij}$.

The $n \times n$ spatial correlation matrix R defines the degree of dependency between design points, based on their observed simulator value. Following MacDonald et al. (2013), we use the Gaussian correlation matrix, R; a special case of the power exponential correlation family defined as

$$R_{ij} = \prod_{k=1}^{d} \exp\{-10^{\beta_k} |x_{ik} - x_{jk}|^{p_k}\} \quad \text{for all } i, j. \tag{1}$$

Here $p_k = 2$ is the smoothness parameter, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is a $1 \times d$ vector of correlation hyper-parameters which measures the sensitivity of the response to the spatial distribution of $|x_{ik} - x_{jk}|^2$ for all $i, j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, d\}$ (Loeppky et al., 2009).

The formulation of the correlation function in Eq. (1) is slightly different than the popular form of Gaussian correlation, which replaces 10^{β_k} with θ_k (e.g., in Ranjan et al., 2011). MacDonald et al. (2013) demonstrate that the deviance surface with β-parametrization shown in Eq. (1) is much easier to optimize as compared to the commonly used θ-parametrization.

Sacks et al. (1989) show that the best linear unbiased predictor (BLUP) at a given point x^* in the input space (typically normalized to $[0, 1]^d$) is

$$\begin{aligned}\hat{y}(x^*) &= \hat{\mu} + r' R^{-1} (Y - \mathbf{1_n}\hat{\mu}) \\ &= \left[\frac{(1 - r' R^{-1} \mathbf{1_n})}{\mathbf{1_n}' R^{-1} \mathbf{1_n}} \mathbf{1_n}' + r' \right] R^{-1} Y \\ &\equiv C' Y,\end{aligned}$$

where $r = [r_1(x^*), \ldots, r_n(x^*)]'$, and $r_i(x^*) = \mathrm{corr}[z(x^*), z(x_i)]$ is the correlation between $z(x^*)$ and $z(x_i)$. The GP model also returns the associated uncertainty estimate, $s^2(x^*)$, as measured by the mean squared error (MSE),

$$s^2(x^*) = E\left[(\hat{y}(x^*) - y(x^*))^2\right] = \hat{\sigma}^2 (1 - 2C'r + C'RC). \tag{2}$$

The model fitting process requires the estimation of μ, σ^2 and β. The closed form estimators of the mean and variance are given by

$$\hat{\mu}(\beta) = (\mathbf{1_n}' R^{-1} \mathbf{1_n})^{-1} (\mathbf{1_n}' R^{-1} Y)$$