Computational Statistics and Data Analysis 71 (2014) 1088–1102

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Optimal design for correlated processes with input-dependent noise

A. Boukouvalas [a,*], D. Cornford [a], M. Stehlík [b]

[a] Non-Linear Complexity Research Group, Aston University, Aston Triangle, Birmingham, United Kingdom

[b] Department of Applied Statistics, Johannes Kepler University in Linz, Austria

ARTICLE INFO

Article history:
Received 14 April 2012
Received in revised form 25 July 2013
Accepted 24 September 2013
Available online 16 October 2013

Keywords:
Optimal design of experiments
Correlated observations
Emulation
Gaussian process
Heteroscedastic noise

ABSTRACT

Optimal design for parameter estimation in Gaussian process regression models with input-dependent noise is examined. The motivation stems from the area of computer experiments, where computationally demanding simulators are approximated using Gaussian process emulators to act as statistical surrogates. In the case of stochastic simulators, which produce a random output for a given set of model inputs, repeated evaluations are useful, supporting the use of replicate observations in the experimental design. The findings are also applicable to the wider context of experimental design for Gaussian process regression and kriging. Designs are proposed with the aim of minimising the variance of the Gaussian process parameter estimates. A heteroscedastic Gaussian process model is presented which allows for an experimental design technique based on an extension of Fisher information to heteroscedastic models. It is empirically shown that the error of the approximation of the parameter variance by the inverse of the Fisher information is reduced as the number of replicated points is increased. Through a series of simulation experiments on both synthetic data and a systems biology stochastic simulator, optimal designs with replicate observations are shown to outperform space-filling designs both with and without replicate observations. Guidance is provided on best practice for optimal experimental design for stochastic response models.

1. Introduction

Design plays an important role in enabling effective fitting and exploitation of a wide variety of statistical models, e.g. regression models such as Gaussian processes. The motivation for this work is a recognition that experimental design plays a crucial part in the building of an emulator (Sacks et al., 1989). The use of emulators, or surrogate statistical representations of computer simulators, provides a solution to the computational constraints that limit a full probabilistic treatment of many simulators. Experimental design is particularly relevant to emulation because we are able to choose the inputs at which the simulator is evaluated with almost complete freedom. The simulator is typically expensive to run, thus it is beneficial to optimise the design given the available *a priori* knowledge.

Most work on emulation has focused on deterministic simulators, where the outputs depend uniquely on the inputs, however it is increasingly common to encounter stochastic simulators, where the randomness is typically associated with interactions which are intrinsically unpredictable or represent some unresolved, essentially random, process within the simulator. Examples of stochastic simulators arise in microsimulation in transport modelling (Rasouli and Timmermans,

[*] Corresponding author. Tel.: +44 1212043922.
E-mail addresses: boukouva@aston.ac.uk, panoramixtb@gmail.com (A. Boukouvalas), D.Cornford@aston.ac.uk (D. Cornford), Milan.Stehlik@jku.at (M. Stehlík).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.024

2012) and biochemical networks of reactions (Wilkinson, 2006). Design and emulation methods developed for deterministic computer experiments need to be extended to be applicable in the stochastic context (Henderson et al., 2009).

A common feature of stochastic simulators is that the variance of the output is input dependent. This requires adaptation of the normal Gaussian Process (GP) regression model (Rasmussen and Williams, 2006). In this paper we introduce a class of heteroscedastic GP models that allow for both flexible variance modelling and tractable calculations for optimal design. Our work extends (Zhu and Stein, 2005) which developed optimal designs for homoscedastic GPs using a Fisher information criterion. This paper expands (Zhu and Stein, 2005) to heteroscedastic GPs with replicated observations. Our approach is general and is relevant to areas such as model-based spatial statistics (Diggle et al., 1998; Stein, 1999), where kriging methods are used, and more general GP regression (Rasmussen and Williams, 2006).

When considering correlated processes, such as GPs, the majority of the results of traditional optimal design, such as the General Equivalence Theorem and the additivity of information matrices do not hold (Müller and Stehlík, 2009). For an overview of classical optimal design theory see Atkinson and Donev (1992) or other standard textbooks. In GP regression, a parametric covariance function is used to model the variance and correlation of the unknown function. The parameters of the covariance are usually estimated by Maximum Likelihood (ML) or Bayesian inference. In this paper, we investigate design under ML estimation, with a focus on best learning the model parameters.

By utilising asymptotic results of estimators, useful approximations to finite sample properties can be constructed. Two asymptotic frameworks are considered in the literature (Zhang and Zimmerman, 2005; Stein, 1999): increasing domain and infill domain asymptotics. It has been found that for certain consistently estimable parameters of exponential kernels with and without a noise term, under ML estimation, approximations corresponding to these two asymptotical frameworks perform about equally well (Zhang and Zimmerman, 2005). For parameters that are not consistently estimable however, the infill asymptotic framework is preferable (Kiselák and Stehlík, 2008). In Mardia and Marshall (1984), it was shown that under increasing domain asymptotics the ML estimator, $\hat{\theta}$, converges in probability to the true parameter, θ, and standard asymptotics hold. Unfortunately no such general results exist under infill asymptotics except for specific classes of covariance functions (Abt and Welch, 1998). A non-asymptotic justification is provided by Pázman (2007) using a truncated function expansion, but is only valid for low process noise levels.

Recently, a 'nearly' universal optimality has been addressed for the case of correlated errors, see e.g. Dette et al. (2013) and references therein, overcoming some of the difficulties in the correlated setup. Exact optimal designs for specific linear models with correlated observations have been investigated (see Kiselák and Stehlík, 2008 and references therein), but even for simple models exact optimal designs are difficult to find.

Optimal design for correlated errors has also been examined under generalised least squares estimation of treatment contrasts in fixed-block effects models where correlation is assumed between treatments within the same block (Uddin, 2008). Within the class of equally replicated designs, designs that minimise the variance of treatment contrasts were found. It was also found that for large positive correlations unequally replicated designs could achieve lower variance values. Although the derivation was only for a specific number of treatments and units, the potential that unequally replicated designs hold for a wider class of scenarios is tantalising and is further investigated in this paper for the GP model case.

Most of the literature on optimal experimental design assumes homoscedastic noise. Optimal design under a fixed basis log-linear-in-parameters model is examined in Tack et al. (2002). Although stochastic processes are not considered, the variance model used is similar to the fixed basis model utilised in this work. They follow a Bayesian approach to design and demonstrate that informative priors lead to more efficient designs.

In certain cases there may exist multiple objective functions which depend upon different information matrices. Compound optimal design provides a general approach, combining multiple such objective functions such as model discrimination (T-optimality) and parameter estimation (A- or D-optimality) via a weighted average of their information matrices (McGree et al., 2008). Compound designs may also be used to generate designs with non-equal emphasis on the trend and covariance parameters (Müller and Stehlík, 2010). Hybrid criteria that explicitly combine prediction and parameter estimation have also been developed (Zimmerman, 2006; Zhu and Stein, 2006). In Zimmerman (2006) such a criterion is defined to minimise the maximum predictive variance as well as a summary of the ML parameter covariance. While this criterion selects observations which reduce parameter uncertainty and predictive uncertainty given the current parameter, it does not take into account the effect of parameter uncertainty on prediction error. To address this issue, Zhu and Stein (2006) propose an amended criterion and derive an iterative algorithm which alternates between optimising the design for covariance estimation and spatial prediction. We note here that a space-filling design does not necessarily minimise the prediction error. For instance if one is interested in optimisation of the Integrated Mean Squared Prediction Error (IMSPE), in one dimension and for an Ornstein–Uhlenbeck process, then the space-filling, i.e. equidistant design, is optimal (Baldi Antognini and Zagoraiou, 2010). However, this property is not generally true in a 2-dimensional design space (Baran et al., 2013). As proven in Baran et al. (2013), a space-filling design does not necessarily reduce the IMSPE more than a design forming a line, which they term monotonic set designs.

Geometric designs such as nested or subsampling designs have been proposed to identify hierarchically related sources of variations. They allow for the estimation of the amount of variation that is derived from each hierarchical level and the determination of the optimal allocation of sampling effort to each level (Green, 1979). Such designs place points at a variety of inter-point distances and may be used for the inference of difficult-to-learn GP correlation length-scale parameters (Pettitt and McBratney, 1993).