Journal of Statistical Planning and Inference 144 (2014) 141–151

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Design of experiments for linear regression models when gradient information is available ☆

Yiou Li *, Fred J. Hickernell

Department of Applied Mathematics, Illinois Institute of Technology, Room 208, Engineering 1 Building, 10 West 32nd Street, Chicago, IL 60616, United States

ARTICLE INFO

Available online 10 February 2013

Keywords:
Gradient information
Computer experiments
Efficiency
Robustness
Integrated mean squared error
Low discrepancy

ABSTRACT

When it is expensive to compute a function value via numerical simulation, obtaining gradient values simultaneously can improve model efficiency. This paper considers the design problem for linear regression models fitted using both function and gradient data. A theoretical upper bound is derived on the scaled integrated mean squared error in terms of the discrepancy of the design, and this bound can be used to choose designs that are both efficient and robust under model uncertainty. Low discrepancy designs are those whose empirical distribution functions match a fixed target distribution. Numerical experiments show that for lower degree polynomial models and for models with many interactions, low discrepancy designs approximating the arcsine distribution are superior. On the other hand, for higher degree polynomial models and those with fewer interactions, low discrepancy designs approximating the uniform distribution are superior.

1. Introduction

For some computer simulations, obtaining a single data point can be computationally expensive, limiting the number of data that one can afford. Thus, the choice of the design, i.e., the set of vectors of input values chosen for running these simulations, is an important problem. In many cases, it is worthwhile to evaluate the gradient of the function relating the inputs to the output using adjoint techniques (Griewank, 2003). This can substantially improve the accuracy of estimating a surrogate model for the computer simulation (Roderick et al., 2010). The gradient of a d-variate function provides d more scalar pieces of information, at a cost of perhaps only several times the cost of a function value alone, irrespective of d, as noted by Roderick et al. (2010). For a large number of parameters (factors) d, the amount of extra information per run is significant.

When d is large and the sample size, n, is modest, low degree polynomial models are often used to approximate the response because of their simple forms. The choices of the polynomial basis and the experimental design both affect the estimation error. For example, for a simplified model of nuclear reactor simulations it was demonstrated by Li et al. (2011) that a better choice of polynomial basis can decrease the estimation error by a factor of 1/10 or better by drastically reducing the condition number of the information matrix.

☆ This research is supported in part by Department of Energy Grant DE-SC0002100.
* Corresponding author. Tel.: +1 3125678980.
E-mail addresses: yli88@hawk.iit.edu (Y. Li), hickernell@iit.edu (F.J. Hickernell).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.01.010

This paper focuses on the choice of the experimental design when gradient information is easily available along with function values. For simplicity, the model misspecification and the numerical or stochastic noise in the observations at two distinct locations are assumed to have negligible correlation. This means that any misspecification or numerical noise has a small length scale compared to the spacing of the design points. The goal is two-fold: to find designs that are *robust* to the lack of knowledge about the model form while at the same time being *efficient*. The statistical model allowing for gradient data is described in Section 2, and an expression for the scaled integrated mean squared error (IMSE) is derived in Section 3. Section 4 considers the situation in which the (polynomial) basis functions used in the regression model are unknown beforehand and must be found by model selection techniques. Our interest is in *low discrepancy designs*. The discrepancy of a design, $\xi=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, is defined in (7), and it measures the difference between the empirical distribution of the design, F_ξ, and a fixed target distribution F_{tar}. Theorem 1 in this paper shows that low discrepancy designs protect against poor IMSE. This result generalizes Theorem 1 of Hickernell and Liu (2002) in two ways, by allowing for gradient data and by allowing the distribution defining the IMSE to differ from the target distribution for the design. Section 5 presents simulation results comparing the IMSE for different models and designs.

The study here is related to that of Morris et al. (1993) who investigated how to predict the response using both function values and gradient information under a Bayesian framework. We assume that the relationship between values of the function and the gradient at different locations is modeled by the mean, which is a linear combination of some polynomial basis, and we assume that any misspecification or noise is modeled by a very spiky covariance kernel. Morris et al. (1993) assumed a constant mean and modeled the relationship of the function and gradient values at different locations through a stationary covariance kernel of product form. Thus, the dependence of the response on the factors is contained in the mean in our setting, whereas it is contained in the covariance kernel in the setting of Morris et al. (1993).

There are several reasons for choosing a spiky covariance kernel. We prefer having a mean that is modeled by a polynomial because polynomials can capture variation of the response on both large and small length scales, and low degree polynomials are manageable even for large d. If one adopts a non-trivial covariance kernel, then there are likely many parameters defining the kernel that require estimation. To avoid the problem of too many kernel parameters, one might adopt a stationary or even radially symmetric kernel. However, we are not convinced that stationarity is necessarily an appropriate assumption for computer experiments, even though it may make sense when the inputs are spatial or temporal variables. Another drawback of radially symmetric kernels or stationary kernels of product form is that they tend to have a vanishing domain of influence as d increases. At the same time, a design with n evenly spread points in a d-dimensional cube of fixed volume has design points separated by a distance proportional to $dn^{-1/d}$ as $d\to\infty$. Thus, for moderate to large d many kernels are effectively spiky anyway. If one tries to avoid the vanishing domain of influence for radially symmetric covariance kernels by making the kernels flat, then one recovers polynomial interpolation (Driscoll and Fornberg, 2002), which is similar to assuming that the mean is modeled by a polynomial. Thus, we do not think that assuming a very spiky kernel is too restrictive, especially for moderate to large d.

Morris et al. (1993) showed that in their setting and in the limiting case of a spiky covariance kernel D-optimality corresponds to maximizing the minimum distance between design points. This maximin criterion is one way to ensure that a design is space filling. In our setting we show the value of designs with low discrepancy, another way of ensuring that a design is space filling.

2. Statistical model

Suppose that an experiment has d variables, and let Ω_j be a measurable set of all possible levels of the jth variable. Common examples of Ω_j are $[-1,1]$ and $\mathbb{R}$. The experimental region, Ω, is some measurable subset of $\Omega_1\times\cdots\times\Omega_d$. An experimental design with n points, $\xi=\{\boldsymbol{x}_i=(x_{i1},\ldots,x_{id}):i=1,\ldots,n\}$, is a subset of Ω with multiple copies of the same point allowed.

Let $\mathcal{H}$ be some vector space of real-valued functions defined on Ω, and let $\mathcal{H}$ include all possible responses to the computer experiment that one might imagine. Define an operator $\mathbf{L}_{\boldsymbol{x}}:\mathcal{H}\to\mathbb{R}^m$, which when applied to a d-variate function $f\in\mathcal{H}$, returns an m-vector of outputs dependent on the point $\boldsymbol{x}\in\Omega$. For traditional experiments, $m=1$, and $\mathbf{L}_{\boldsymbol{x}}$ is the scalar evaluation functional, i.e., $\mathbf{L}_{\boldsymbol{x}}f=f(\boldsymbol{x})$. For the applications that motivate this study, $m=d+1$, and $\mathbf{L}_{\boldsymbol{x}}$ returns both the function and gradient values at $\boldsymbol{x}$

$$\mathbf{L}_{\boldsymbol{x}}f=\left(f(\boldsymbol{x}),\frac{\partial f}{\partial x_1}(\boldsymbol{x}),\ldots,\frac{\partial f}{\partial x_d}(\boldsymbol{x})\right)^T.$$

One can imagine cases where $\mathbf{L}_{\boldsymbol{x}}$ includes higher order derivatives as well. For a vector function $\boldsymbol{f}=(f_1,\ldots,f_\ell)^T:\Omega\to\mathbb{R}^\ell$, the definition of this operator is extended to give an $m\times\ell$ matrix: $\mathbf{L}_{\boldsymbol{x}}\boldsymbol{f}^T=(\mathbf{L}_{\boldsymbol{x}}f_1,\ldots,\mathbf{L}_{\boldsymbol{x}}f_\ell)$.

A linear regression model can be written as

$$\tilde{\boldsymbol{y}}_i=(\mathbf{L}_{\boldsymbol{x}_i}\boldsymbol{g}^T)\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}}_i,\quad i=1,\ldots,n, \tag{1}$$

where $\tilde{\boldsymbol{y}}_i$ denotes the observed $m\times 1$ vector response at the design point $\boldsymbol{x}_i$, $\boldsymbol{g}=(g_1,\ldots,g_k)^T$ is the vector of basis functions (polynomials), $\boldsymbol{\beta}$ is the $k\times 1$ regression coefficient to be estimated, and $\tilde{\boldsymbol{\varepsilon}}_i$ is the error in estimating the response by the linear combination of k basis functions. It is assumed for simplicity that the m-vectors $\tilde{\boldsymbol{\varepsilon}}_1,\ldots,\tilde{\boldsymbol{\varepsilon}}_n$ are independent and identically distributed with zero mean and covariance matrix $\sigma^2\tilde{\Lambda}$. Here, σ^2 is the overall magnitude of the covariance