Journal of Statistical Planning and Inference 144 (2014) 63–68

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Maximin distance optimal designs for computer experiments with time-varying inputs and outputs

Max D. Morris [a,b,*]

[a] Department of Statistics, Iowa State University, United States
[b] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, United States

ARTICLE INFO

Available online 27 September 2012

Keywords:
Computer experiment
Dynamic model
Gaussian stochastic process
Maximin distance design
Meta-model
Surrogate

ABSTRACT

Computer models of dynamic systems produce outputs that are functions of time; models that solve systems of differential equations often have this character. Time-indexed inputs, such as the functions that describe time-varying boundary conditions, are also common with such models. Morris (2012) described a generalization of the Gaussian process often used to produce "meta-models" when inputs are finite-dimensional vectors, that can be used in the functional input setting, and showed how the maximin distance design optimality criterion (Johnson et al., 1990) can also be extended to this case. This paper describes an upper bound on the maximin distance criterion for functional inputs. A class of designs that are optimal under certain conditions is also presented; while these designs are of limited practical value, they show that the derived bound cannot be improved in the general case.

1. Introduction and notation

The statistical treatment of *computer experiments*, in which a *computer model* of a physical system is the object of experimental focus, has received substantial attention in the statistical literature over the last two decades. The computer model is often an implementation of a complicated deterministic function for which evaluation requires the specification of quantities that complete the definition of a problem to be solved. Because the function is complex, it is often difficult or impossible to understand analytically, and so it is implemented as a computer model $y = f(x)$, where x represents the variables for which values must be specified, or *inputs*; f is the computer model; and y represents the variables that constitute the problem solution, or *outputs*. f is generally also complex and its evaluation may require substantial computational effort. A common final or intermediate goal of a computer experiment is the development of a *surrogate* or *meta-model*, an approximation of the computer model based on a finite number of executions of f, that can be used to quickly predict what the outputs would be at new values of the inputs along with useful quantification of the uncertainty associated with these predictions. A few of the broadly cited papers on this topic are Sacks et al. (1989), Currin et al. (1991), Kennedy and O'Hagan (2001), and Higdon et al. (2004). In each of these papers, and many others, a Gaussian stochastic process (GaSP) is used as the statistical basis for formulating a surrogate for predicting a scalar- or vector-valued y from a scalar- or vector-valued x.

In many applications, the computer models of greatest interest are written to mimic dynamic physical processes; the literature describing methodology based on GaSP meta-models includes examples involving DNA population dynamics

* Correspondence address: Department of Statistics, Iowa State University Ames, IA 50014, United States. Tel.: +1 515 291 2775; fax: +1 515 294 4040.
E-mail address: mmorris@iastate.edu

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.09.005

(e.g. Henderson et al., 2010), the design of helicopter rotor blades (Booker et al., 1999), and the global climate (e.g. Tebaldi et al., 2005). These computer models are often based on systems of ordinary or partial differential equations that require specification of time-varying quantities as inputs, e.g. boundary conditions or so-called "forcing functions", and their solutions are outputs that are also functions of time (and sometimes also of space). Smooth functional outputs can often be reduced to vectors of low dimension for practical analysis through, for example, principle components representation, e.g. Higdon et al. (2008).

Morris (2012) introduced a generalization of the GaSP meta-model for the case in which the inputs x are expressed as a function of time t, along with a corresponding generalization of the maximin distance design criterion (Johnson et al., 1990), along the following lines. An experimental design is a collection of n functional inputs, $D=\{x_1(t),x_2(t),\ldots,x_n(t)\}$, defined for $t\in[0,T]$. The GaSP treats the responses at a particular time s that result from different input functions x_i and x_j as correlated with $Corr(y_i(s),y_j(s))=\exp(-\theta d_{ij}^s)$, where d_{ij}^s is called the *correlation distance* between x_i and x_j over $[0,s]$, but treats responses at any two distinct times as independent whether they are generated by the same input function or not. One way to look at this is to say that the correlation distance associated with any two responses at different times is infinite. This modeling assumption is admittedly non-intuitive, since functional output that is continuous in time suggests dependence across time, but incorporating such dependence is not straight-forward in this case. For models that have vector-valued inputs and functional outputs, temporal correlations *can* be incorporated in a simple way through the *separable* or *product* correlation structure, with one factor reflecting temporal correlation over the time interval (s_1,s_2), and the remaining factors expressing correlation associated with the inputs used in modeling output at either t_1 or t_2. However, in our context $Corr(y_i(s_1),y_j(s_1))\neq Corr(y_i(s_2),y_j(s_2))$, preventing the use of this approach. As a practical matter, the assumption of temporal independence may be less critical than it initially seems. For example, where input-time product correlations *are* appropriate and observed data include outputs at each of a collection of times S for selected inputs, predictive interpolation for any time in S depends only on data observed at that time, i.e. data at other times are not additionally informative. (See, for example, Drignei and Morris, 2006.)

Generalizing a commonly used product correlation model described by Sacks et al. (1989) for vector-valued inputs, the correlation distance between inputs x_i and x_j through time s is defined as $d_{ij}^s=\int_0^s w_s(s-u)|x_i(u)-x_j(u)|^\alpha\,du$ for a specified $\alpha\in(0,2]$ and w_s defined over $[0,s]$. A popular choice is $\alpha=2$, the so-called "Gaussian" correlation form, which is particularly appropriate for modeling outputs that are smooth functions of inputs, a condition that is generally true for models that implement systems of partial differential equations. The positive-valued *weight functions* $w_s(-)$, one defined for each time for which output is to be predicted, generalize the individual weights associated with each dimension of the usual vector-input model. Effective specification of w_s requires at least some understanding of the computer model under study. For example, for systems in which outputs are more dependent on recent inputs than those in the more distant past, reasonable weight functions generally take larger values for small arguments (recent times) and smaller values for larger arguments; the example of Section 3.1 features a very simple weight function of this type. In contrast, for systems that are sensitive to early inputs but become less influenced by inputs as time progresses, larger values of w_s may be associated with larger values of its argument (more distant times).

Suppose we want to predict outputs over a set of times S. The maximin (Mm) distance criterion for D over output times in S, under a model of independence between outputs at any two distinct times, is

$$\phi=\min_{1\le i<j\le n,s\in S} d_{i,j}^s \tag{1}$$

and a Mm distance design for a given value of n is any D for which ϕ is maximized. Morris (2012) constructed a Mm distance design numerically for an experiment involving a dynamic model of the reaction of stem cell "kinetics" to exposure to ionizing radiation. In Section 2 of this paper, we derive a general upper bound for ϕ for even values of n. In Section 3, we present a class of designs for $n=0[\mathrm{mod}\ 4]$ that achieve this bound. The designs in this class are of limited practical value, but serve to show that the bound of Section 2 cannot be improved in the general case.

2. An upper bound for ϕ

Suppose for simplicity that the problem is scaled so that $T=1$, and that each input function is constrained so that $0\le x\le 1$, the latter being required so that distance is constrained. Further, require that for each $s\in S$

$$\int_0^s w_s(s-u)\,du=1 \tag{2}$$

so that distances $d_{i,j}^s$ are comparable across all $s\in S$.

Result. Restrict n to be an even integer. Then an upper bound on ϕ (as defined in Eq. (1)) is $\frac{1}{2}n/(n-1)$.

Proof. Define the *total* distance between the elements of a design at any time s to be

$$\psi^s=\sum_{i<j} d_{i,j}^s.$$