Journal of Statistical Planning and Inference 154 (2014) 3–11

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Optimal design for smooth supersaturated models

Ron A. Bates [a], Peter R. Curtis [b], Hugo Maruri-Aguilar [b,*], Henry P. Wynn [c]

[a] Rolls Royce plc, PO Box 31, ML-80, Derby DE24 8BJ, UK
[b] School of Mathematical Sciences, Queen Mary University, Mile End E1 4NS, UK
[c] Department of Statistics, London School of Economics, London WC2A 2AE, UK

ARTICLE INFO

Available online 1 December 2013

Keywords:
Splines
Kernel smoothing
Experimental design
Algebraic statistics

ABSTRACT

Smooth supersaturated models are interpolation models in which the underlying model size, and typically the degree, is higher than would normally be used in statistics, but where the extra degrees of freedom are used to make the model smooth using a standard second derivative measure of smoothness. Here, the solution is derived from a closed-form quadratic programme, leading to tractable matrix representations. This representation aids considerably in the choice of optimal knots in the interpolation case and in the optimal design when the SSM is used as a way of obtaining kernels, but where the statistical problem is set up separately. Some examples are given in one and two dimensions.

1. Introduction

The basic idea of smooth supersaturated models (SSM) on which this paper is founded appears in Bates et al. (in press), and follows a few years of development (an arXiv version has been available since 2009), particularly in the context of computer experiments. In the present paper a theory of optimal experimental designs for SSM is developed. Insofar as a high order SSM can be considered as an approximation to a multidimensional spline, a solution to the optimal design problem for SSM gives an approximate solution to optimal design for splines which, in high dimensions is not very much studied, but see Studden and VanArmann (1969) and Dette et al. (2001) for some work in the area.

As with splines there is the problem of the choice of knots. We shall explain how optimal design and optimal choice of knots arise as two different problems and suggest solutions to each.

Let $(x_1, \ldots, x_k)$ be a general point in R^k. A monomial is defined by a non-negative integer vector $\alpha = (\alpha_1, \ldots, \alpha_k)$:

$$x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_k}.$$

Following the experimental design avenue of algebraic statistics (Pistone et al., 2001), it is clear that for observations over any design D_n with n points in R^k there is at least one exact polynomial interpolator. Specifically, let a design be defined as a set of distinct points $D = \{x^{(1)}, \ldots, x^{(n)}\}$ in R^k. A general polynomial model can be written as

$$\eta(x) = \sum_{\alpha \in M} \theta_\alpha x^\alpha, \tag{1}$$

for some set M of distinct index vectors, α. The algebraic theory shows that there is always a set of indices M for which we have an exact interpolation of a set of observations $y = \{y_1, \ldots, y_n\}$ at design points $x^{(1)}, \ldots, x^{(n)}$ respectively and for which the

* Corresponding author.
E-mail address: H.Maruri-Aguilar@qmul.ac.uk (H. Maruri-Aguilar).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.11.014

size of M is n: $|M| = n$. Moreover there is a method of finding M based on Gröbner bases and M has a hierarchical structure: if $\alpha \in M$ then $\beta \in M$, for any $\beta \leq \alpha$, where $\leq$ is the usual entrywise ordering. We speak informally of "the model M". The algebraic method is the starting point or at least a theoretical underpinning for SSM.

A supersaturated polynomial model is one in which the number of parameters, p, is larger than that suggested by the size of the design, the number of observations n. In the present day terminology we may say that this is a "p bigger than n problem". However, the SSM approach is a little different. Initially we increase the size of the model, $|M|$ so that $|M| > p$. In statistical terminology this leaves $|M| - n$ "free" degrees of freedom which we use to increase the smoothness of the model, in a well defined sense, while still interpolating the original data set y.

2. The SSM construction

We start with a data set y over a design D_n, with n points (D for short). The data y is given as a column vector of size n. Write the vector of model terms as $f(x) = (x^\alpha : \alpha \in M)^T$ so that

$$\eta(x) = f(x)^T \theta, \tag{2}$$

where θ is the vector of coefficients for monomials in $f(x)$ in a suitable order according to elements of M. Denote the number of model terms as $|M| = N$ and assume that $N > n$. Let the region of interest be $\Omega \subset R^k$, which we call the "integration region". Our measure of smoothness Bates et al. (2003) is

$$\phi(M, \Omega) = \int_{\Omega} 1 \leq \sum_{i,j \leq k} \left(\frac{\partial^2 \eta(x)}{\partial x_i \partial x_j} \right)^2 dx.$$

The smooth supersaturated model given by $\{y, M, D, \Omega\}$ is $\eta(x)$ with θ chosen to solve the optimisation problem:

$$\min \phi(M, \Omega) \quad \text{subject to } \eta(x^{(i)}) = y_i, \quad i = 1, \ldots, n. \tag{3}$$

In short, the SSM is a maximally smooth interpolator. A key observation is that this problem can be written as a constrained quadratic optimisation problem and therefore has a closed form solution. We summarise the results of Bates et al. (in press). First write

$$K = \int_{\Omega} 1 \leq \sum_{i,j \leq k} f^{(i,j)^T} f^{(i,j)} \, dx, \tag{4}$$

where $f^{(i,j)} = \partial^2 f(x) / \partial x_i \partial x_j$. The matrix K is symmetric of size N, whose elements are roughness measures between pairs of monomials in $f(x)$.

Example 1. In one dimension ($k=1$), the hierarchical basis with N elements is $1, x, x^2, \ldots, x^{N-1}$, i.e. $M = \{0, 1, \ldots, N-1\}$. If the integration region is $\Omega = [0, 1]$ then for $0 \leq i,\ j \leq N-1$, the entry $K_{i+1 j+1}$ of K is $(i^2 - i)(j^2 - j)/(i + j - 3)$ if $i + j \neq 3$ and zero otherwise. When $N=6$, then K is of size six, with the first two rows and columns being equal to zero, and the lower right block is

$$\begin{pmatrix} 4 & 6 & 8 & 10 \\ 6 & 12 & 18 & 24 \\ 8 & 18 & 144/5 & 40 \\ 10 & 24 & 40 & 400/7 \end{pmatrix}.$$

Example 2. Consider $M = \{(0, 0), (0, 1), (1, 0), (0, 2), (1, 1), (1, 2)\}$ for $k=2$, i.e. the model has terms $1, x_2, x_1, x_2^2, x_1 x_2, x_1 x_2^2$. For $\Omega = [0, 1]^2$, the matrix K has the first three rows and columns equal to zero, and lower right block is

$$\begin{pmatrix} 4 & 0 & 2 \\ 0 & 2 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

Let the design matrix for the model given by M and the design D be

$$X = \{x^\alpha\}_{x \in D,\ \alpha \in M}.$$

The n rows of X are indexed by the design points in D and the N columns by the model monomials of M. This is the familiar supersaturated design matrix which has more columns than rows. We shall assume that X has full rank, n. The choice of M to guarantee this is discussed at some length in Bates et al. (in press) and it is here where the methods of algebraic statistics are useful as a guide. For example, we may just add model terms to a model basis with n terms which we already know, from the algebra, is full rank.