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*

Abstract

Smoooth 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 somple examples are given in one and two dimensions.

Key words: splines, kernel smoothing, experimental design, algebraic statistics.

1 Introduction

The basic idea of smooth supersaturated models (SSM) on which this paper is founded appears in [2], 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. In so far 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

* Corresponding author.
Email address: `H.Maruri-Aguilar@qmul.ac.uk` (Hugo Maruri-Aguilar).

22 November 2013

which, in high dimensions is not very much studied: but see [11,4] 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 knots 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 [9], 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. Specificaly, 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 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 which the number of parameters, p, is larger than is suggested by the size of the design, the number of observations n. In the present day terminology we may say 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 is

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

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

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

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 [2]. First write

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

where $f^{(i,j)} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. 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 $0 \le i, j \le N-1$, the entry $K_{i+1,j+1}$ of K is $(i^2-i)(j^2-j)/(i+j-3)$ if $i + j \ne 3$ 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 $1, x_2, x_1, x_2^2, x_1x_2, x_1x_2^2$. For $\Omega = [0, 1]^2$, the matrix K has the first three rows and columns equal to zero, and lower right block

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

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 [2] 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.

The optimisation problem (3) can be stated as minimisation of $\theta^T K \theta$ subject to $X\theta = y$. This constrained problem is equivalent to the system

$$\begin{pmatrix} X & 0_{n,n} \\ K & -X^T \end{pmatrix} \begin{pmatrix} \theta \\ \lambda \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \tag{5}$$

where λ is a vector of Lagrange multipliers, see [2]. The following lemma states the form of the inverse matrix required to solve the problem.

Lemma 3 *Let C be the $(n+N) \times (n+N)$ matrix in the left of Equation (5). This matrix C is created by concatenation of X, K and a zero matrix of size $n \times n$. Assume that C is invertible. Then*

(1) its inverse C^{-1} can be written in the following block form

$$C^{-1} = \begin{pmatrix} H & P \\ Q & -H^T \end{pmatrix}, \tag{6}$$

where P and Q are symmetric matrices of sizes N and n respectively and H is a matrix of size $N \times n$.

(2) the matrices inside C^{-1} and C satisfy the following relations: $XH = I_n$, $XP = 0_{n,N}$, $KH = X^T Q$ and $PK + HX = I_N$, where I_n, I_N are identity matrices of sizes n and N.

The optimum θ is given by

$$\theta^* = Hy. \tag{7}$$

The optimal (minimum) value of ϕ is given by the quadratic form in the data

$$\phi^* = y^t Q y.$$

Recall that the matrix Q is the bottom left $n \times n$ submatrix of the inverse matrix C^{-1} above. An equivalent expression for Q shows that it is symmetric and positive semidefinite:

$$Q = H^T K H.$$

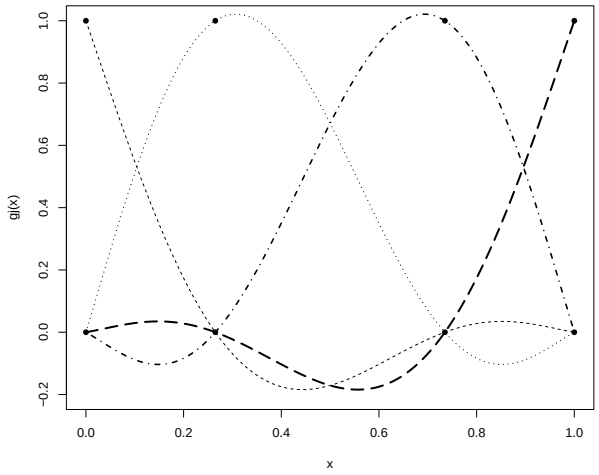

Fig. 1. Smooth kernels $g_1(x), \ldots, g_4(x)$ of Example 4.

2.1 The fit: smooth kernels

From the linearity of the θ^* in the data y we can write the fit

$$\hat{\eta}(x) = f(x)^T \theta^* = \sum_{j=1}^{n} g_j(x) y_j, \tag{8}$$

where the polynomials $g_j(x),\ j = 1, \ldots n$, have the indicator property at the design points:

$$g_j(x^{(i)}) = \delta_{i,j} \ \text{ for } \ i, j = 1, \ldots, n. \tag{9}$$

Note that the indicator property is equivalent to $XH = I_n$ above. We refer to the polynomials $g_j(x)$ as *smooth kernels* and they are the elementary building blocks for the smooth model, i.e. the fit in (8) is a linear combination of kernels, where the coefficients are the data values y_j.

We find the kernels $g_j(x)$ by substituting y for unit basis vectors e_j in the formula for $\hat{\eta}(x)$:

$$g_j(x) = f(x)^T \theta_j^*,$$

where

$$\theta_j^* = H e_j \tag{10}$$

for $j = 1, \ldots, n$. That is, the vector θ_j^* is the j-th column of matrix H defined above, and thus H of Equation (6) has the following form:

$$H = (\, \theta_1^* : \theta_2^* : \cdots : \theta_n^* \,). \tag{11}$$

Example 4 Consider the one-dimensional model with $N = 6$ of Example 1, and design $D = \{0, 0.265, 0.735, 1\}$. Four smooth kernels of degree five were constructed; they are depicted in Figure 1 together with design points to show that they satisfy the indicator property of Equation (9).

We emphasize that these kernels depend on the original design D via the full design model matrix X. But the terminology of splines encourages us to call the points in D *knots*. The construction of kernels does not depend on the data y. We can use the knots simply as an ingredient in constructing kernels. We are then free to take observations at any other points and fit a model using the basis vector $g(x) = (g_1(x), \ldots, g_n(x))^T$. For clarity, we refer to the original design D as the *knot design*, which leaves us free to use the term *design*, labeled d, for the points at which actual observations are taken. It may be that the knot design and the design are the same, but this is not essential and typically not optimal. Indeed we will separate out the optimal knot problem from the optimal design problem, and this is a main heuristic of this paper.

3 Large bases and splines

The optimality criteria ϕ is precisely that for which splines, particularly cubic beta splines and thin plate splines are known to be optimal in the class of functions with bounded, continuous derivatives of given order. It ought to be true that as the size of the basis given by $|M|$ gets large in an appropriate way, we tend to splines. We state a somewhat more general definition than is available in the literature and refer to [6].

Definition 5 *Let Ω be a bounded Lebesgue integrable region in R^d, let $D = \{x^{(1)}, \ldots, x^{(n)}\}$ be a knot design, and let y be a set of observations at D. Then a generalised Duchon spline given $\{\Omega, D, y\}$ is a function which achieves*

$$\inf \phi(s(x)) = \phi^*(\Omega, D, y)$$

over all twice continuously differentiable functions $s(x)$ which interpolate y over D.

We need a basic approximation theorem, which is adapted below without further exposition from the Sobolev representation theorem [6].

Lemma 6 *Let $s(x)$ be a twice continuously differentiable function over a bounded integrable region $\Omega \subset R^d$. Let $s(x)$ interpolate values y over D. Then given $\delta > 0$ there exists a multivariate polynomial $p(x)$, also interpolating y over D such that, for all $x \in \Omega$,*

$$|p(x) - s(x)| < \delta.$$

Define the degree of a monomial α as $|\alpha| = \sum_{i=1}^{d} \alpha_i$. We need one more property, which is defined for an index set M.

Definition 7 *The degree order $r(M)$ of an index set M with associated basis $\{x^\alpha \in M\}$ is the minimum value of r such that M contains all α with $|\alpha| \leq r$.*

Our main theorem is as follows.

Theorem 1 *Let ϕ_M be the value of criterion ϕ for an SSM built with model basis $\{x^\alpha,\ \alpha \in M\}$ in a bounded integrable region Ω with a knot design D and data y. Then for any nested sequence of models $M_s, s = 1, \ldots$ with $r(M_k) \to \infty$, the quantity ϕ_M converges to $\phi^*(\Omega, D, y)$ of the Duchon spline.*

Proof. Let $M_1 \subset M_2 \subset \cdots$ be a nested sequence of M such that $r(M_j) \to \infty$ with j. Let $s_j(x), j = 1, \ldots$, be a sequence of twice continuously differentiable function interpolating y over D with $\phi(s_j(x)) \downarrow \phi^*(\Omega, D, y)$. Fix j, let $p_j(x)$ be the approximating polynomial to $s_j(x)$ according to Lemma 6, and let r_j be the maximal degree of polynomials in the set $\{p_k(x) : j = 1, \ldots, j\}$. By the definition of degree order there will be an integer $k(j)$ such that $r(M_{k(j)}) \geq r_j$.

Note that $r > s$ implies $\phi_{M_r} \leq \phi_{M_s}$, because for SSM the optimisation problem based on M_r has more degrees of freedom that that based on M_s. We thus, given $\delta > 0$, have the inequalities

$$\phi^* \leq \phi_{M_q} \leq \phi_{M_{k(j)}} \leq \phi(p_j(x)) \leq \phi(s_j(x)) + \delta,$$

for any $q \geq k(j)$. Now letting $\delta \to 0$ and using the convergence of $s_j(x)$, we are done. □

4 Orthogonal polynomials and computation

The computations required for smooth supersaturated models can be easily implemented when the number of factors is small, say less than ten, and also the number of observations and terms is not huge, say in the region of a few hundred observations and a few hundred extra terms. However, as the number of factors increase, computations may slow considerably. Computing the matrix K involves summation over k^2 pairs of factors, and in each case a matrix of size $N \times N$ is to be computed. Some efficiency can be attained by noting that only $k + \binom{k}{2}$ pairs are needed, but even for a moderate number of factors such as $k = 15$ and if the model involves, say, $N = 200$ terms, computing the matrix K requires summing $10 + \binom{15}{2} = 115$ matrices of size 200×200. Among other issues is the inversion of matrix C, which can be prone to numerical instability. Therefore there is pressing need for efficient and numerically stable computations.

Building the models using orthogonal polynomials instead of pure monomials is appropriate when the region for interpolation and the integration region Ω

are the same. When $\Omega = [-1,1]^k$, or $\Omega = [0,1]^k$, the orthogonal polynomials with respect to uniform measure are the appropriate Legendre polynomials, and we will study the $\Omega = [-1,1]^k$ case here. We will discuss briefly the more general case in the last section; [12] is the classical text.

The Legendre-based SSM model is

$$\eta_L(x) = \sum_{\alpha \in M} \gamma_\alpha L_\alpha(x), \tag{12}$$

where γ_α is the parameter for term $L_\alpha(x)$, defined as $L_\alpha(x) := \prod_{i=1}^k L_{\alpha_i}(x_i)$ where $L_{\alpha_i}(x_i)$ is the usual Legendre polynomial. The first few Legendre polynomials are $L_0(x) = 1, L_1(x) = x, L_2(x) = (3x^2-1)/2$ and $L_3(x) = (5x^3-3x)/2$. They are ortogonal over $[-1,1]$.

Due to the hierarchical structure of M, the Legendre-based SSM is linearly related to the general formulation of Equation (2). Consider the index set $M = \{0,1,2,3\}$, then $\eta_L(x) = (1, x, (3x^2-1)/2, (5x^3-3x)/2)\gamma = (1, x, x^2, x^3)A\gamma$ where the matrix A contains in each column the coefficients of the Legendre polynomials, i.e.

$$A = \begin{pmatrix} 1 & 0 & -1/2 & 0 \\ 0 & 1 & 0 & -3/2 \\ 0 & 0 & 3/2 & 0 \\ 0 & 0 & 0 & 5/2 \end{pmatrix}.$$

It is a direct consequence that the models (2) and (12) are linearly related, and their parameters are linked in general through the matrix A as $\theta = A\gamma$.

As a more complicated example, a monomial term such as $x_1x_2^2x_4$ with index $(1,2,0,4)$ will be replaced by $L_1(x_1)L_2(x_2)L_0(x_3)L_1(x_4) = x_1(3x_2^2-1)x_4/2 = 3x_1x_2^2x_4/2 - x_1x_4/2$. Given that the set M is hierarchical, if term $x_1x_2^2x_4$ is in the model, so is term x_1x_4 and therefore no extra terms apart from those already in M appear in the model.

Computations of the K matrix are also simplified when using the Legendre basis, attaining matrices with a more sparse structure. Also in our experience, the matrices obtained are better conditioned.

Example 8 For the same M as in Example 1 and region $\Omega = [-1,1]$, the

lower right block of K is

$$\begin{pmatrix} 18 & 0 & 60 & 0 \\ 0 & 150 & 0 & 420 \\ 60 & 0 & 690 & 0 \\ 0 & 420 & 0 & 2310 \end{pmatrix}.$$

The condition number for the non singular submatrix of K above is 188.3. This is a considerable reduction compared with that of the submatrix shown in Example 1 which is 8467.2.

The following theorem states the general case for the construction of K. We omit the proof, which is by direct construction.

Theorem 2 *Let M be the set of indexes of a multivariate basis of Legendre polynomial products $g(\mathbf{x})^T = \{L_\alpha(\mathbf{x}) : \alpha \in M\}$. Let $m, n \in 1, \ldots, N$ index pairs of elements in M; let $i, j \in 1, \ldots, k$ index pairs of indeterminates and let $\alpha^i = \min(\alpha_{mi}, \alpha_{ni})$, $\alpha^I = \max(\alpha_{mi}, \alpha_{ni})$, and $\alpha^j = \min(\alpha_{mj}, \alpha_{nj})$. Then*

(1) the entry (m, n) of the matrix $\int_\Omega {g^{(i,j)}}^T g^{(i,j)} dx$ with $\Omega = [-1, 1]^k$ is given by $g_{mn}(i, j)$ as follows:

(i) $g_{mn}(i, j) =$

$$\frac{1}{24}\alpha^i\left(\alpha^i - 1\right)\left(\alpha^i + 1\right)\left(\alpha^i + 2\right)\left(3{\alpha^I}^2 + 3\alpha^I + 6 - {\alpha^i}^2 - \alpha^i\right)\prod_{r \neq i}\frac{2\delta_{\alpha_{mr}\alpha_{nr}}}{2\alpha_{mr} + 1},$$

if α_{mi} and α_{ni} are both even or both odd and $i = j$,

(ii) $g_{mn}(i, j) = \alpha^i\alpha^j\left(\alpha^i + 1\right)\left(\alpha^j + 1\right)\prod_{r \neq i,j}\frac{2\delta_{\alpha_{mr}\alpha_{nr}}}{2\alpha_{mr}+1}$,
if $i \neq j$, and α_{mi}, α_{ni} are both odd or both even, and α_{mj}, α_{nj} are both odd or both even, or

(iii) $g_{mn}(i, j) = 0$, otherwise.

(2) The entry (m, n) of the K matrix is

(i) zero if three or more entries of α_m, α_n are pairwise different,
(ii) $g_{mn}(i, j)$ if α_m, α_n differ in two entries,
(iii) $\sum_{l=1}^k g_{mn}(i, l)$ if α_m, α_n differ in only one entry and
(iv) $\sum_{i,j} g_{mn}(i, j)$ if $m = n$, i.e. $\alpha_m = \alpha_n$.

In the theorem above, $\delta_{\alpha_{mr}\alpha_{nr}}$ is the delta function that equals one if $\alpha_{mr} = \alpha_{nr}$ and zero otherwise.

Example 9 Consider the basis of Example 8. We compute the element $3, 5$ of K so $m = 3, n = 5$. Being a univariate model, $i = j = 1$ so that $\alpha_m = \alpha_3 = 2$, $\alpha_n = \alpha_5 = 4$ and thus $\alpha^i = 2, \alpha^I = 4$ and $\alpha^j = 2$. Since α_m and α_n have only

one element and $k = 1$ the entry is given by $g_{mn}(i, j)$. We compute this using the case 1(i) above since α_m and α_n are both even and $i = j = 1$ thus

$$g_{mn}(i, j) = \frac{2}{24}(2 - 1)(2 + 1)(2 + 2)(3 \cdot 4^2 + 3 \cdot 4 + 6 - 4 - 2) = 60.$$

Example 10 As a more complex case, consider the multivariate basis with $N = 27$ elements and highest term with exponent $(2, 2, 2)$, i.e. the basis is $M = \{(r, s, t) : 0 \leq r, s, t \leq 2\}$. The matrix K has $N^2 = 729$ entries and here we give an example of entries of K showing the sparsity achieved by the use of Legendre polynomials. The parity condition in the definition of $g_{mn}(i, j)$ results in every non-diagonal entry of K being zero since the only possible pair α_{mi}, α_{ni} that fulfils the condition of having the same parity and being different is $(0, 2)$. Hence $\alpha^i = 0$ and $g_{mn}(i, j) = 0$ for the first two cases in the definition. We need to see indices of degree 3 or more before non-zero off-diagonal entries appear.

Example 11 Now consider the multivariate basis with $N = 64$ elements and highest term with exponent $(3, 3, 3)$, i.e. the basis is $M = \{(r, s, t) : 0 \leq r, s, t \leq 3\}$. The matrix K has $N^2 = 4096$ entries. The entry $(K)_{m,n}$ corresponding to a crossing of terms with indices that differ on three or more entries such as $(1, 1, 0)$ and $(0, 2, 2)$ is zero as stated in the first case above; this first case induces 1728 entries of K to be zero. There are 1728 entries corresponding to crossed terms with exactly two different indices entries, such as $(1, 2, 3)$ and $(1, 0, 2)$ and these vanish if any of the four differing indices are 0 or if there is a parity difference between either index, because of the definition of $g_{mn}(i, j)$. This induces a further 1680 zeroes. Should the crossed terms differ in only one index entry, the corresponding entry will disappear if one of the differing indices is zero or they have different parity. Hence 486 more entries vanish. Finally, any diagonal entry will vanish if the associated term is the constant term or linear in only one variate, so in this example that is another 4 entries. In total we have 3898 zero entries in this K matrix out of a possible 4096.

5 Optimal knots

If one considers the interpolation discussed here as a method which can be applied to any data set y, at the selected knot design D, then there is a sense in which D should be independent of the actual (true) underlying process yielding the data.

For a model given by M and given that $\phi^* = y^T Q y$, we may consider simple measures on the matrix Q. Consider, for motivation the case when the vector of observations y has a multivariate distribution with mean vector μ

and covariance matrix Σ. Then the expected roughness of the SSM model is $E(\phi^*) = E(y^TQy) = \text{trace}(Q\Sigma) + \mu^TQ\mu$. The second term depends on the unknown mean and matrix Q. If we consider uncorrelated observation with equal variance σ^2 and zero mean we have $\Sigma = \sigma^2 I$ and then

$$E(\phi^*) = \sigma^2 \text{trace}(Q).$$

Several generic criteria suggest themselves at this point. If $\lambda_1(Q) \geq \cdots \geq \lambda_n(Q) \geq 0$ are the ordered eigenvalues of Q, possible criteria are

$$\begin{aligned}
\psi_1(Q) &= \lambda_1(Q), \\
\psi_2(Q) &= \text{trace}(Q) = \sum_i \lambda_i(Q) \text{ and} \\
\psi_3(Q) &= \prod_{j:\lambda_j(Q)>0} \lambda_j(Q).
\end{aligned}$$

In each case we would seek to minimise the criteria over the choice of the knot design D for a fixed model M. Note that each kernel $g_j(x)$ is optimally smooth under the kernel restriction given by (9). Thus, we can interpret criteria ψ_2 above as minimising the average smoothness of the kernels.

Example 12 Consider a univariate model with terms $1, x, x^2, \ldots, x^{N-1}$ and the criterion $\psi_1(Q)$ which we use to create a design with $n = 4$ points. This implies that the number of model terms has to satisfy $N \geq 5$, i.e. the smallest model has monomials up to degree four. Analytic computation of eigenvalues of $Q = H^TKH$ is possible, and we searched for knot designs of four points using values of N up to 26 and restricting the knots to be inside $\Omega = [0, 1]$. For $N = 5$, knots are $\{0, 0.2351, 0.7649, 1\}$, and when $N = 6$ we obtain the knots which we used for the kernels in Example 4. Figure 2 depicts the knot designs obtained against values of N; the right panel shows a close-up of the left plot. In the close-up an interesting feature appears, which consists of knots being equal for some adjacent values of N. For $N > 20$, knots seem to converge to a limiting value and for $N = 26$ the knots are $\{0, 0.2732, 0.7268, 1\}$.

Example 13 A bivariate supersaturated model with $N = 25$ terms was built, consisting of all monomials which divide the term $x_1^4x_2^4$, i.e. the exponent set was $M = \{(i, j) : 0 \leq i, j \leq 4\}$. This model was used to search for a Q-optimal set of $n = 16$ knots in the region $\Omega = [0, 1]^2$, using criterion $\psi_1(Q)$. Although numerical search of the knots in general position in Ω is possible, we searched for a product set of knots of the form $D = \{0, a, 1-a, 1\} \otimes \{0, a, 1-a, 1\}$ and numerical optimization yielded the value $a = 0.2304$. When the supersaturated model was enlarged to be of size $N = 36$, using all monomials that divide $x_1^5x_2^5$, the Q-optimal design with a product structure as above was found with $a = 0.2689$.

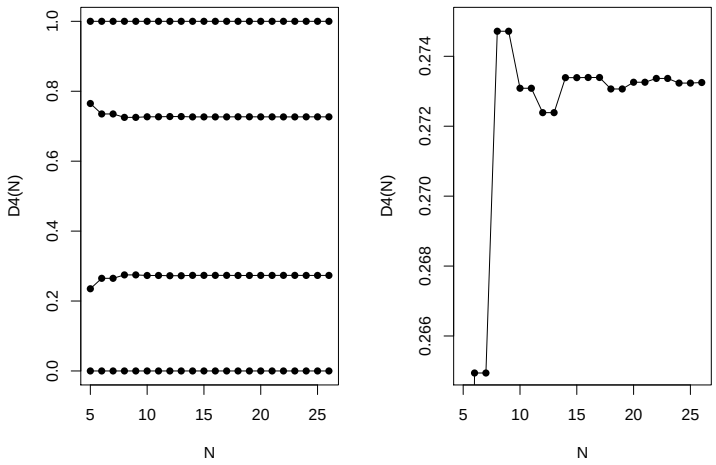

Fig. 2. Q-optimal knots (left) and detail of the plot (right), see of Example 12.

6 Optimal design

As discussed briefly above we shall study the pure optimal design problem for the kernel basis given by the SSM basis: $g(x) = (g_1(x), \ldots, g_n(x))^T$. For this we take the classical approach. We assume that we have a design d with m points: $d = \{z^{(i)}, i = 1, \ldots, m\}$. The model for observation Y_i, taken at point $z^{(i)}$, is

$$Y_i = g(z^{(i)})^T \beta + \epsilon_i.$$

Here β is the vector of parameters and errors ϵ_i are independent and normally distributed with zero mean and variance σ^2. To study design we need the design matrix

$$Z = \left(g_j(z^{(i)}) \right).$$

But from (9) and for $j = 1, \ldots, n$ we have

$$g_j(z^{(i)}) = f(z^{(i)})^T \theta_j^*,$$

therefore for optimal design we need to consider

$$Z^T Z = H^T X_d^T X_d H, \tag{13}$$

where X_d is design matrix with the model $f(x)$ but with the design d and $H = (\theta_1^* : \cdots : \theta_n^*)$ as defined in (11) with the θ_j^* given by (10). Note that H depends on the choice of knot design.

The form of (13), makes it, as expected, straightforward to import some optimal design theory. In the continuous design theory of Kiefer and Wolwitz [7], a discrete design d is replaced by a design measure ξ over a design space $\mathcal{X}$ and the matrix $X^T X$ by a moment matrix

$$\mathcal{M}(\xi) = \int_{\mathcal{X}} f(x) f(x)^T \xi(dx).$$

The optimal design theory then relies heavily on the fact that over the class of moment matrices many of the best known optimal design criteria are convex. But because of the nice form (13) we have

$$H^T \mathcal{M}(\xi) H,$$

where $\mathcal{M}(\xi)$ is the moment matrix for the measure extension of the design d.

Thus we have D_Q-optimality, defined as

$$\max \det(H^T \mathcal{M}(\xi) H^T).$$

This criterion has a general equivalence theorem which is obtained by writing down the General Equivalence Theorem (GET) of Kiefer and Wolfowitz for the model with $g(x) = (g_1(x), \ldots, g_n(x))$.

Theorem 3 *The following are equivalent for a design measure ξ^* on $\mathcal{X}$*

(1) ξ^* is D_Q-optimal,
(2) ξ^* achieves: $\min_\xi \max_{x \in \mathcal{X}} \; d_Q(x, \xi)$,
(3) $\max_{x \in \mathcal{X}} \; d_Q(x, \xi^*) = n$,

where

$$d_Q(x, \xi) = f(x)^T H (H^T \mathcal{M}(\xi) H)^{-1} H^T f(x)$$

is the (generalised) variance function for the smooth kernel model based on $g(x) = (g_1(x), \ldots, g_n(x))$.

The optimal measure theory embodied in Theorem 3 is a special case of a range of duality theorems for convex functional in measure, or moment space $\{\mathcal{M}(\xi)\}$, familiar from optimal design theory and given general expression in the book [10]. Theorem 3 may be the first to incorporate smoothness into the optimal design criteria. However, rather than develop the measure-based results in greater detail we simply give some examples of the same calibre as the introductory and the optimal knot examples. These are computed numerically but part (3) of Theorem 3 is used to verify optimality to a substantial degree of accuracy.

Example 14 Here we continue with the set of $n = 16$ knots for the basis with $N = 25$ terms of Example 13. An exact D_Q-optimal design of $m = 16$ runs was built for the linear model with kernels $g_1(x), \ldots, g_{16}(x)$ and design region $\mathcal{X} = [0, 1]^2$. Numerical search yielded the D_Q-optimal design d^*, which was in general position in $\mathcal{X}$. The left panel of Figure 3 shows countour plot for the standardized variance function $d_Q(x, d^*)/n$, together with points in the knot design D (squares) and optimal design d^* (circles). Recall that the knots were chosen to lie in a grid, see Example 13. The design points in d^* are virtually in a grid configuration; indeed points in d^* are close to points in

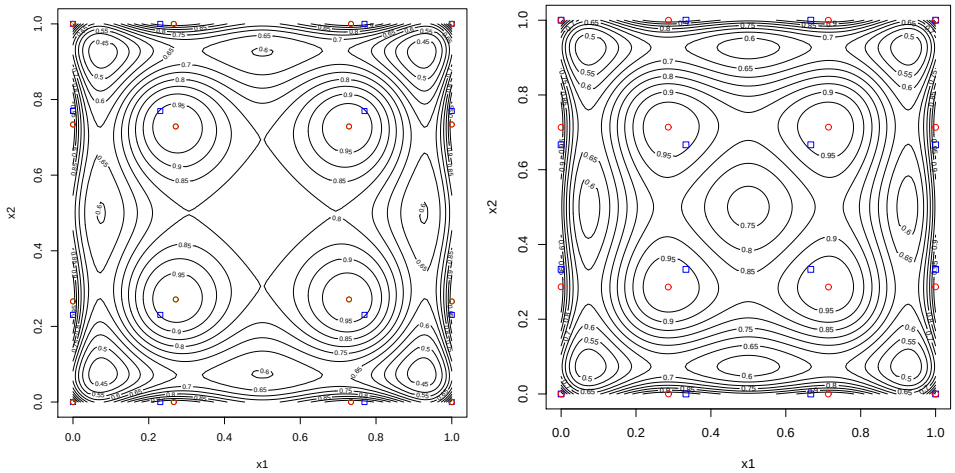

Fig. 3. Standardized variance function contours and knot design D (squares) and optimal design d^* (circles) for Example 14 (left) and Example 15 (right).

grid $\tilde{d} = \{0, 0.2687, 0.7313, 1\}^2$. The efficiency of $\tilde{d}$ relative to d^* was 99.52% and standardized efficiency of 99.97%.

Example 15 It is possible to construct D-optimal designs for kernels starting from an arbitrary set of knots. Using the same extended basis of Example 14 and a knot design with grid structure $D = \{0, 1/3, 2/3, 1\}^2$, a D-optimal design was built for the kernels $g_j(x)$. The search this time yielded a grid design $d^* = \{0, 0.286, 0.714, 1\}^2$. The right panel in Figure 3 shows countours for the standardized variance function, together with points in D (squares) and d^* (circles). Note how relative to the left panel in the same Figure, the variance function is slightly lower in the central region which suggests a flatter central region for the D_Q-optimal design of Example 14.

7 Discussion

An issue not covered in this paper is that fact that the integration region can be quite general (provided the integral exists) and one could also change the measure integration of Equation (4) all without changing the basic theory. This is implied by also by the general definition of a Duchon spline given in Section 3. We could also have given a version for general measures. Thus one can see the paper as an approach to solving spline-like problems of a quite general nature, approximately. By problems we mean: fitting functions, optimal knot designs and optimal statistical designs. Clearly it has not been possible to cover all region, all models M and all criteria, but we trust that we have indicated the possibilities.

The use of orthogonal polynomials we feel is important both theoretically

and computationally. It is of interest that the use of orthogonal polynomials in Sobolev spaces, motivated by the theory of splines, is an active branch of approximation theory. So, theoretically, one can consider our use if them, combined with large $|M| = N$ as remaining in the space of traditional orthogonal polynomials but using them for models which may be close to, or a sub-class of, a more general space. It is straightforward to define orthogonal polynomials in terms of moments for a general measure in R^k, and therefore, for example, for uniform measure over some non-standard region.

Although there are recent papers on Lasso and other regularized methods for hierarchical models, such as [3], for the $p > n$ case and extensive work on regularization for non-parametric regression, we believe that the SSM methods which can be interpreted as a type of smoothness regularization deserve a place for complex modelling over complex regions, partly because of the tractability of using polynomials. It is notable that the approach of working with polynomials as building material to create kernels with special properties is already being used in signal processing, see [8].

Acknowledgments

The third and fourth authors acknowledge support under EPSRC UK grants EP/D048893/1 and EP/K036106/1 and EPSRC PhD funding for the first author. The fourth author acknowledges a Leverhulme emeritus fellowship.

References

[1] Bates, R., Giglio, B., and Wynn, H. (2003). A global selection procedure for polynomial interpolators. *Technometrics*, 45(3):246–255.

[2] Bates, R., Maruri-Aguilar, H. and Wynn, H.P. (2013). Smooth supersaturated models. *Journal of Statistical Computation and Simulation*, in press.

[3] Bien, J., Taylor, J. and Tibshirani, R. (2013). A lasso for hierarchical interactions. Annals of Statistics, 41:111-1141.

[4] Dette, H., Melas, V, B. and Pepelyshev, A. (2011). Optimal design for smoothing splines. *Annals of the Institute of Statistical Mathematics*, 63:981-1003.

[5] Duchon, J. (1976). *Splines minimizing rotation invariant semi-norms in Sobolev spaces*. Lectures notes in Math. 571, Springer, Berlin.

[6] Dupont, T. and Scott, P. (1980). Polynomial approximation of functions in Sobolev spaces. *Mathematics of Computation* 34:441-463.

[7] Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Annals of Mathematical Statistics*, 30(2): 271–294.

[8] Lebrun, L. and Selensnick, I. (2004). Gröbner bases and wavelet design. *Journal of Symbolic Computation* 37:227-259.

[9] Pistone, G., Riccomagno, E., and Wynn, H. P. (2001). *Algebraic Statistics*, volume 89 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton.

[10] Pukelsheim, P. (1993). *Optimal design of experiments*. Wiley, New York.

[11] Studden, W.J. and VanArmann, D. J. (1969). Admissible designs for spline polynomial spline regression Annals of Mathematical Statistics, 40:1557-1569.

[12] Szego, G. (1939). *Orthogonal Polynomials*. Colloqium publications XXIII American Mathematical Society.