Working Paper M10/02
Methodology

Optimal Design For Additive

Partially Nonlinear Models

Stefanie Biedermann, Holger Dette, David C. Woods

Abstract

We develop optimal design theory for additive partially nonlinear regression models, and show that D-optimal designs can be found as the products of the corresponding D-optimal designs in one dimension. For partially nonlinear models, D-optimal designs depend on the unknown nonlinear model parameters, and misspeci_cations of these parameters can lead to poor designs. Hence we generalise our results to parameter robust optimality criteria, namely Bayesian and standardised maximin D-optimality. A su_cient condition under which analogous results hold for Ds-optimality is derived to accommodate situations in which only a subset of the model parameters is of interest. To facilitate prediction of the response at unobserved locations, we prove similar results for Q-optimality in the class of all product designs. The usefulness of this approach is demonstrated through an application from the automotive industry where optimal designs for least squares regression splines are determined and compared with designs commonly used in practice.

Optimal design for additive partially nonlinear models

Stefanie Biedermann

University of Southampton

School of Mathematics

Southampton SO17 1BJ

UK

email: S.Biedermann@soton.ac.uk

Holger Dette

Ruhr-Universität Bochum

Fakultät für Mathematik

44780 Bochum

Germany

email: Holger.Dette@ruhr-uni-bochum.de

David C. Woods

University of Southampton

Statistical Sciences Research Institute

Southampton SO17 1BJ

UK

email: D.Woods@soton.ac.uk

April 8, 2010

Abstract

We develop optimal design theory for additive partially nonlinear regression models, and show that D-optimal designs can be found as the products of the corresponding D-optimal designs in one dimension. For partially nonlinear models, D-optimal designs depend on the unknown nonlinear model parameters, and misspecifications of these parameters can lead to poor designs. Hence we generalise our results to parameter robust optimality criteria, namely Bayesian and standardised maximin D-optimality. A sufficient condition under which analogous results hold for D_s-optimality is derived to accommodate situations in which only a subset of the model parameters is of interest. To facilitate prediction of the response at unobserved locations, we prove similar results for Q-optimality in the class of all product designs. The usefulness of this approach is demonstrated through an application from the automotive industry where optimal designs for least squares regression splines are determined and compared with designs commonly used in practice.

Keywords: Additive model, Bayesian D-optimality, partially nonlinear model, product design, Q-optimality, standardised maximin D-optimality.

1   Introduction

In many real life problems, regression models are used to describe the relationship between a real valued response and a high dimensional predictor. Because of the so called curse of dimensionality, a popular strategy in this situation is to additively combine univariate basis functions for different variables for dimensionality reduction (see [3] or [13] among others). These models are parsimonious, but sufficiently flexible to capture the important features as long as the variables do not interact. A typical class of such models in K variables $x_1, \ldots, x_K$ is defined by

$$\mu(x, \tau) \; = \; \theta_0 + \sum_{k=1}^{K} \tilde{\mu}_k(x_k, \tau_k), \quad x \; = \; (x_1, \ldots, x_K)^T, \tag{1}$$

where the kth regression function $\tilde{\mu}_k$ depends only on the variable $x_k \in \chi_k \subset \mathbb{R}$. We assume that n observations

$$Y_i \; = \; \mu(x_{(i)}, \tau) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

at experimental conditions $x_{(1)}, \ldots x_{(n)}$ are available, where $\tau = (\theta_0, \tau_1^T, \ldots, \tau_K^T)^T$ is the vector of unknown parameters and the errors ε_i, $i = 1, \ldots, n$, are independent and identically distributed according to a $\mathcal{N}(0, \sigma^2)$ distribution. Model (2) is partially nonlinear in its parameter vector τ if the vector $\tau = (\theta^T, \lambda^T)^T$ can be split into a linear component θ and a nonlinear component λ such that the Fisher information at the point x can be represented as

$$M(x, \theta, \lambda) \; = \; C_\theta f(x, \lambda) f^T(x, \lambda) C_\theta^T. \tag{3}$$

Here C_θ is a nonsingular square matrix depending only on the linear parameters θ, but neither on λ nor x, and $f(x, \lambda)$ is a vector of functions depending on x and the nonlinear parameters λ only (see e.g. [14] and [15]).

Partially nonlinear models are widely used in various application areas. They often have the form

$$\tilde{\mu}_k(x_k, \tau) = \sum_{i=1}^{l_k} \theta_{k,i} h_{k,i}(x_k, \lambda_k)$$

for the different additive components $\tilde{\mu}_k$, where $h_{k,i}(z, \lambda)$ are linearly independent functions ($i = 1, \ldots, l_k$). This class of models is extremely flexible and contains exponential models used in toxicokinetic and pharmacokinetic experiments (see, e.g., [2] or [1]), rational models ([8]) and logarithmic models. A particularly popular subclass of the partially nonlinear models, which in

2

fact motivated this research, are least squares regression splines leading to a model of the form (1) where the individual components $\tilde{\mu}_k$ are given by

$$\tilde{\mu}_k(x_k, \tau_k) = \sum_{i=1}^{l_k} \theta_{k,i} x_k^i + \sum_{i=1}^{r_k} \sum_{j=0}^{l_{k,i}-1} \theta_{k,i,j} (x_k - \lambda_{k,i})_+^{m_k-j}.$$

Here, the knots $\lambda_{k,i}$ are assumed to be unknown and thus require estimation. Due to their conceptual simplicity combined with their high flexibility ([9]) these models are widely used in applications such as engine-mapping experiments from the automotive industry ([12]), dynamic programming, computer models and chromatography ([5], [20], [10] and [16]).

The present paper is devoted to the problem of constructing efficient designs for partially non-linear regression models with multivariable predictors. While optimal design has been discussed intensively for multivariable linear models (see e.g. [17] or [19]), much less effort has been devoted to develop theory for finding efficient designs in the nonlinear case. [18] investigate locally D-optimal designs for linear heteroscedastic models, whereas [11] consider locally D-optimal designs for multivariable generalised linear models. Many commonly applied models are not covered by this literature, and no theoretical results for multivariable models have yet been provided that take into account parameter uncertainty which arises naturally from nonlinear models. The goal of the present paper is to present a unified approach for characterising robust designs for a large class of multivariable nonlinear models. In particular, we investigate the relationship between optimal designs in the additive model and the corresponding single variable models with respect to several local and robust criteria. In many cases the optimal designs for the multivariable case can be constructed from the univariate optimal designs for the single variable models, which are considerably easier to calculate analytically and numerically.

2 Optimal design for partially nonlinear models

Recall the definition of the additive model in (1) and consider the projection

$$\mu_k(x_k, \tau_k) = \theta_{k,0} + \tilde{\mu}_k(x_k, \tau_k) \tag{4}$$

onto its kth variable, where $\theta_{k,0}$ are intercept terms, $k = 1, \ldots, K$. We assume that $\mu_k(x_k, \tau_k)$ is a partially nonlinear regression model in one variable $x_k \in \chi_k \subset \mathbb{R}$ with parameter vector $\tau_k = (\theta_k^T, \lambda_k^T)^T$, which means that its Fisher information matrix at the point x_k has the form

$$M_k(x_k, \theta_k, \lambda_k) = C_{\theta_k} f_k(x_k, \lambda_k) f_k^T(x_k, \lambda_k) C_{\theta_k}^T.$$

It is obvious that the additive model (1) is then also partially nonlinear in the sense of (3).

If a single variable model contains additive terms, which are only distinguishable through different values of the nonlinear parameters in λ_k, we need to restrict the possible values for the components of λ_k to avoid identifiability problems. For example, for a spline featuring additive

terms $\theta_{k,i}(x_k - \lambda_{k,i})_+^l$ and $\theta_{k,j}(x_k - \lambda_{k,j})_+^l$ for $i \neq j$ the case $\lambda_{k,i} = \lambda_{k,j}$ must be excluded. In particular when specifying prior distributions for the nonlinear parameters λ_k, see later, we must ensure that unidentifiable parameter combinations will not be in the support of the priors.

An approximate design $\xi = \{x_{(1)}, \ldots, x_{(m)}; w_1, \ldots, w_m\}$ is a probability measure with finite support on the design space $\chi = \chi_1 \times \ldots \times \chi_K \subset \mathbb{R}^K$, i.e. $x_{(i)} \in \chi$, $i = 1, \ldots, m$. The observations are taken at the support points of the design, and the number of observations in each point $x_{(i)}$ is proportional to the weight w_i.

Let $g_k(x_k, \theta_k, \lambda_k) = C_{\theta_k} f_k(x_k, \lambda_k)$ and $g(x, \theta, \lambda) = C_\theta f(x, \lambda)$ be the respective vectors of parameter sensitivities for the kth single variable model (4), $k = 1, \ldots, K$, and the additive model (1). The Fisher information of a design ξ for the additive model (1) is then given by the matrix

$$M(\xi, \theta, \lambda) = \sum_{i=1}^m w_i \, g(x_{(i)}, \theta, \lambda) g^T(x_{(i)}, \theta, \lambda) = C_\theta I(\xi, \lambda) C_\theta^T \tag{5}$$

where $I(\xi, \lambda) = \sum_{i=1}^m w_i f(x_{(i)}, \lambda) f^T(x_{(i)}, \lambda)$. Using properties of the determinant, it follows that

$$|M(\xi, \theta, \lambda)| = |C_\theta I(\xi, \lambda) C_\theta^T| = |C_\theta|^2 \, |I(\xi, \lambda)|,$$

so the same design $\xi_{D,\lambda}^*$ will maximise the determinants of the Fisher information $M(\xi, \theta, \lambda)$ and of the matrix $I(\xi, \lambda)$, which is independent of θ and which will be denoted as an information matrix in what follows. The design $\xi_{D,\lambda}^*$ will only depend on the vector of the unknown nonlinear parameters λ, but not on the linear parameters θ. Following [6], we call a design $\xi_{D,\lambda}^*$ locally D-optimal if it maximises the determinant of the Fisher information matrix for given λ, i.e.

$$\xi_{D,\lambda}^* = \arg \max_\xi |I(\xi, \lambda)|.$$

3 D- and D_s-optimal designs

3.1 Locally and robust D-optimal designs

The concept of local D-optimality requires knowledge of the unknown parameter vector λ. If λ is misspecified at the design stage, the design may be inefficient. Several approaches to overcome the parameter dependency of optimal designs in nonlinear models have been suggested. We will focus on two non-sequential concepts: Bayesian D-optimality (see, e.g., [4]) and standardised maximin D-optimality ([7]).

When some prior knowledge about the nonlinear parameters is available, which can be summarised in a prior distribution π on the parameter space Λ, it is reasonable to use a Bayesian optimality criterion which averages the original criterion over the plausible values for λ. The Bayesian D-optimality criterion function with respect to the prior π on Λ is given by

$$\Phi_{D,\pi}(\xi) = \int_\Lambda \log |I(\xi, \lambda)| \, d\pi(\lambda), \tag{6}$$

4

and is maximised with respect to the design ξ.

Alternatively, the problem of specifying a prior on the knots can be avoided by using a maximin approach guarding the experiment against the worst case scenario. This is a more cautious approach than the Bayesian, and is recommended in the absence of adequate prior knowledge. The standardised maximin D-optimality criterion is defined as maximisation of

$$\Psi_{D,\Lambda}(\xi) = \inf_{\lambda \in \Lambda} \frac{|I(\xi, \lambda)|}{|I(\xi_{D,\lambda}^*, \lambda)|}, \tag{7}$$

where $\xi_{D,\lambda}^*$ is the locally D-optimal design with respect to λ. Throughout this paper we assume that $\Lambda = \Lambda_1 \times \ldots \times \Lambda_K$, where $\Lambda_k \subset \mathbb{R}^{r_k}$, $k = 1, \ldots, K$, are sets of plausible values for the parameters $\lambda_k = (\lambda_{k,1}, \ldots, \lambda_{k,r_k})^T$, specified by the experimenter, which exclude parameter combinations that lead to identifiability problems in the additive model (1). The following result states how Bayesian and standardised maximin D-optimal designs for the additive model (1) can be constructed from the corresponding Bayesian and standardised maximin D-optimal designs for the single variable models (4).

Theorem 1

(a) Let $\xi_{D,\pi_1}^*, \ldots, \xi_{D,\pi_K}^*$ be the respective Bayesian D-optimal designs for the single variable models (4) with respect to the priors π_k, and let π be the product prior $\pi_1 \otimes \ldots \otimes \pi_K$, $k = 1, \ldots, K$. Then the product design $\xi_{D,\pi}^* = \xi_{D,\pi_1}^* \otimes \xi_{D,\pi_2}^* \otimes \ldots \otimes \xi_{D,\pi_K}^*$ is Bayesian D-optimal with respect to π for the additive model (1).

(b) Let $\xi_{D,\Lambda_1}^*, \ldots, \xi_{D,\Lambda_K}^*$ be the standardised maximin D-optimal designs with respect to the compact sets Λ_k, $k = 1, \ldots, K$, for the single variable models (4). Then the product design $\xi_{D,\Lambda}^* = \xi_{D,\Lambda_1}^* \otimes \xi_{D,\Lambda_2}^* \otimes \ldots \otimes \xi_{D,\Lambda_K}^*$ is standardised maximin D-optimal with respect to Λ for the additive model (1).

See Appendix A.1 for the proof of Theorem 1. Local D-optimality can be viewed as a special case of Bayesian or standardized maximin D-optimality, where the set Λ is a singleton.

Remark 1 *The number of support points of product designs quickly increases in higher dimensions, so the experimenter may prefer to run a smaller design. It is still vital to have an optimal design as a benchmark to compare competing designs against to avoid inefficient designs being run, which could result in unreliable conclusions from the data.*

From Corollary 5.4 in [19] we obtain that a necessary condition for local D-optimality of a design ξ in the additive model (1) is local D-optimality of the marginals of ξ in the corresponding single variable models (4). If the D-optimal designs ξ_{D,λ_k}^*, $k = 1, \ldots, K$, for the single variable models are unique, any D-optimal design for the additive model must therefore have its support contained in the support of the product design $\xi_{D,\lambda_1}^* \otimes \ldots \otimes \xi_{D,\lambda_K}^*$. A numerical determination of a locally D-optimal design with possibly smaller support than the product design can therefore be restricted to the support of the product design.

3.2 Application - Engine mapping experiment

The purpose of engine mapping experiments as considered in [12] is to model a measure of engine performance as a function of several adjustable engine variables. The data for such an experiment described in [21] give rise to an additive spline model for the maximum brake torque timing of an engine in the three variables "speed", "load" and "air-fuel ratio". The corresponding single variable models are the cubic spline model

$$\mu_1(x_1, \tau_1) = \theta_{1,0} + \theta_{1,1}x_1 + \theta_{1,2}x_1^2 + \theta_{1,3}x_1^3 + \theta_{1,4}(x_1 - \lambda_{1,1})_+^3 \,, \tag{8}$$

with unknown knot $\lambda_{1,1}$ for the variable "speed", and quadratic polynomials for "load" and "air-fuel ratio", respectively. [12] use a complicated numerical search algorithm in three variables to find optimal designs for this model, which could have been simplified considerably if our results had been available to them.

We use the engine mapping model to demonstrate the usefulness of designed experiments and to address the issue of robustness in the situation where the knot location is unknown. The data imply that the knot $\lambda_{1,1}$ should be in the interval $[0, 0.6]$. We investigate the performance of:

(1) ξ_1, the locally D-optimal design for the midpoint $\lambda_{1,1} = 0.3$

(2) ξ_2, the Bayesian D-optimal design with respect to π_1, the uniform distribution on the seven points $0, 0.1, \ldots, 0.6$

(3) ξ_3 the Bayesian D-optimal design with respect to the continuous uniform prior on $[0, 0.6]$ with (6) approximated using π_2, a uniform distribution on 121 equidistant points from 0 to 0.6

(4) ξ_4, the product of: the uniform design on 11 equidistant points in $[-1, 1]$ for the splined variable, and two uniform designs on $\{-1, 0, 1\}$ for the other two variables

(5) ξ_5, the product of: an irregularly spaced uniform design with 11 points, where the points are more concentrated in the interval for the knot, for the splined variable, and two uniform designs on $\{-1, 0, 1\}$ for the other two variables.

The two latter designs are commonly applied in such experiments. Note that designs $\xi_1 - \xi_3$ are also product designs where the marginals for the variables "load" and "air-fuel ratio" are the same as for ξ_4 and ξ_5. To compare designs, we define the relative D-efficiency of a design ξ_r compared with a design ξ_s as

$$\text{eff}_{rel,D}(\xi_r, \xi_s, \lambda) = \left(\frac{|I(\xi_r, \lambda)|}{|I(\xi_s, \lambda)|} \right)^{1/p},$$

where p is the number of model parameters. Designs $\xi_1 - \xi_3$ were calculated numerically, and the marginals of ξ_1, ξ_2, ξ_4 and ξ_5 for the splined variable "speed" are depicted in the left part
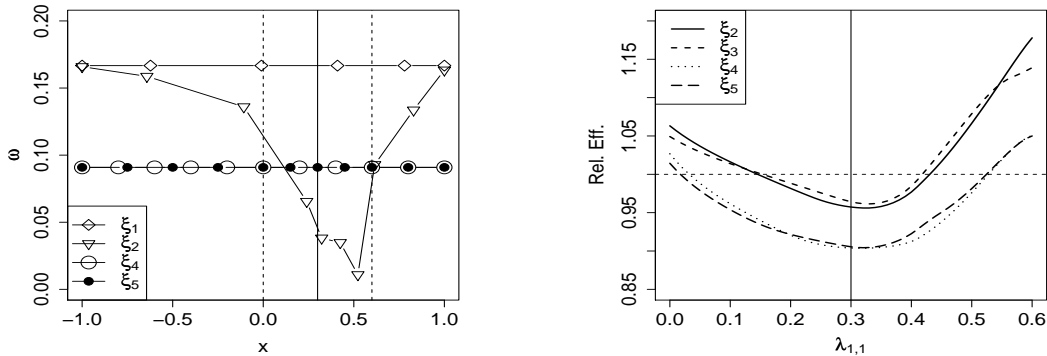
Figure 1: *Left: Support points and weights of the marginals of ξ_1, ξ_2, ξ_4 and ξ_5 for the splined variable "speed", described by (8). Right: Relative D-efficiencies of designs $\xi_2 - \xi_5$ compared with ξ_1, plotted against the possible knots in the interval $[0, 0.6]$.*

of Figure 1. Design ξ_3 is very similar to ξ_2 and therefore not shown. Figure 1 also shows the relative D-efficiencies of $\xi_2 - \xi_5$ compared with ξ_1.

Figure 1 shows that the commonly used designs ξ_4 and ξ_5 are uniformly inferior to the Bayesian designs. The spacing of the support points for the uniform designs seems to have little effect. We also considered two similarly structured uniform designs with 21 support points, which were both outperformed by the 11-point designs and are therefore not shown here. For those designs, unequal spacing appeared to have an adverse effect on efficiency. We further note that in the interval from about 0.14 to 0.42 the Bayesian D-optimal designs are slightly less efficient, but outperform the locally D-optimal design if the knot is closer to the boundary. Both Bayesian designs have similar relative D-efficiencies, with the Bayesian D-optimal design with respect to π_1 having slightly higher efficiency around the boundary, and the Bayesian D-optimal design with respect to π_2 being somewhat more efficient in the interior. For a large degree of uncertainty it is thus recommended to use a Bayesian D-optimal design, where a close approximation to the continuous uniform prior does not seem to have an advantage over a prior distribution with a relatively crude space-filling support on the same interval.

3.3 Optimal designs for estimating subsets of the model parameters

In some practical problems, the experimenter's main interest is in estimating a subset of the parameters only. For example, for spline models interest may be in the knot locations, since they indicate at which experimental conditions the behaviour of the regression function changes and thus provide insight into the complexity of the response. In other examples, the intercept may be less important than the parameters describing the shape of the response curve. In what follows, we therefore investigate D_s-optimal designs for the estimation of subsets $\varphi \subset (\theta^T, \lambda^T)^T$ of the parameters. This means we minimise the determinant of the asymptotic covariance matrix

7

for the estimator of φ, or equivalently, maximise the function

$$\psi_s(M(\xi,\theta,\lambda)) = |(A^T M^-(\xi,\theta,\lambda)A)^{-1}|. \tag{9}$$

The matrix $A^T = (J_s \mid 0_{s\times(p-s)})$ consists of two blocks, where J_s is the identity matrix of size s where s is the number of parameters in φ and $0_{s\times(p-s)}$ is a zero matrix of size $s \times (p-s)$. Without loss of generality, throughout this section we have re-ordered the rows and columns of the information matrix $M(\xi,\theta,\lambda)$ such that the top left corner of size $s \times s$ of this matrix corresponds to the derivatives of the regression function with respect to φ, and also re-ordered the rows and columns of $I(\xi,\lambda)$ accordingly, i.e. through multiplication with the appropriate permutation matrix P from the left and its transpose from the right. Here, $M^-(\xi,\theta,\lambda)$ and $I^-(\xi,\lambda)$ denote the respective generalised inverses of the matrices $M(\xi,\theta,\lambda)$ and $I(\xi,\lambda)$. The design ξ must ensure that the parameters in φ are estimable, i.e. the matrix $A^T M^-(\xi,\theta,\lambda)A$ must be non-singular.

Lemma 1 shows that for many interesting problems we can restrict ourselves to considering the simpler problem of maximising $\psi_s(I(\xi,\lambda)) = |(A^T I^-(\xi,\lambda)A)^{-1}|$, and that consequently D_s-optimal designs for estimating φ in model (1) do not depend on the linear model parameters θ. The proof of Lemma 1 can be found in Appendix A.2.

Lemma 1 *Let $\tilde{C}_\theta = PC_\theta P^T$, and suppose the lower left block of size $(p-s) \times s$ of $\tilde{C}_\theta$, $\tilde{C}_{\theta,21}$, is the zero matrix $0_{(p-s)\times s}$. Then there exists a positive constant c_θ, depending only on θ but neither on λ nor on the design ξ, such that*

$$|(A^T M^-(\xi,\theta,\lambda)A)^{-1}| = c_\theta|(A^T I^-(\xi,\lambda)A)^{-1}|.$$

Remark 2 *For many partially nonlinear models, the condition on $\tilde{C}_\theta$ is satisfied for all subsets φ. For example, if the marginal models are of the form $\mu_k(x_k,\tau_k) = \theta_{k,0} + \sum_{i=1}^{l_k} \theta_{k,i} h_{k,i}(x_k,\lambda_{k,i})$, $k = 1,\ldots,K$, the matrix C_θ is diagonal, and so is $\tilde{C}_\theta = PC_\theta P^T$. For different models it will depend on the subset of interest, φ, if the condition is satisfied. For spline models the condition is met, for example, for $\varphi = (\theta_\varphi^T, \lambda^T)^T$ where θ_φ can be any subset of θ, or for $\varphi = \lambda_\varphi$ where λ_φ can be any subset of λ.*

We now consider Bayesian and standardised maximin D_s-optimality, where a Bayesian D_s-optimal design with respect to a prior π on Λ maximises

$$\Phi_{D_s,\pi}(\xi) = \int_\Lambda \log \psi_s(I(\xi,\lambda))\, d\pi(\lambda),$$

and a standardised maximin D_s-optimal design with respect to Λ maximises

$$\Psi_{D_s,\Lambda}(\xi) = \inf_{\lambda\in\Lambda} \frac{\psi_s(I(\xi,\lambda))}{\psi_s(I(\xi^*_{D_s,\lambda},\lambda))}.$$

Here $\xi^*_{D_s,\lambda}$ denotes the locally D_s-optimal design with respect to λ. Analogous to Section 3.1, we show that under certain conditions the product of designs which are Bayesian (standardised

maximin) D_s-optimal for estimating the set of parameters φ_k in the kth marginal model (4) are Bayesian (standardised maximin) D_s-optimal for estimating the set $\varphi = (\varphi_1^T, \ldots, \varphi_K^T)^T$ in the additive model (1). Local D_s-optimality is embedded in this result as the special case where the set Λ is a singleton. The proof of Theorem 2 is in Appendix A.3.

Theorem 2 *Suppose that $\tilde{C}_{\theta,21} = 0_{(p-s) \times s}$, and that the subset $\varphi = (\varphi_1^T, \ldots, \varphi_K^T)^T$ of parameters of interest does not contain the intercept.*

(a) *Let π be a product prior for $\lambda \in \Lambda$ with marginals π_k, and let ξ_{D_s,π_k}^* denote the Bayesian D_s-optimal design for estimating φ_k with respect to π_k in the single variable models, $k = 1, \ldots, K$. Then the product design $\xi_{D_s,\pi}^* = \xi_{D_s,\pi_1}^* \otimes \xi_{D_s,\pi_2}^* \otimes \ldots \otimes \xi_{D_s,\pi_K}^*$ is Bayesian D_s-optimal for estimating φ with respect to π in the additive model (1).*

(b) *Let ξ_{D_s,Λ_k}^* be the standardised maximin D_s-optimal design for estimating φ_k with respect to Λ_k, $k = 1, \ldots, K$, in the single variable models (4) for compact parameter spaces Λ_k. Then the product design $\xi_{D_s,\Lambda}^* = \xi_{D_s,\Lambda_1}^* \otimes \xi_{D_s,\Lambda_2}^* \otimes \ldots \otimes \xi_{D_s,\Lambda_K}^*$ is standardised maximin D_s-optimal for estimating φ with respect to the parameter space Λ in the additive model (1).*

4 Optimal designs for prediction of the response surface

The experimenter may be more interested in the prediction of the response surface at different points than in the particular values of the unknown parameters. Spline models, for example, are mainly used for prediction rather than estimation. A first order approximation to the variance of $\hat{\mu}(x, \tau)$ at some point $x = (x_1, \ldots, x_K) \in \mathbb{R}^K$ is given by

$$Var\left(\hat{\mu}(x, \tau)\right) = g^T(x, \theta, \lambda) M^{-1}(\xi, \theta, \lambda) g(x, \theta, \lambda) = f^T(x, \lambda) I^{-1}(\xi, \lambda) f(x, \lambda).$$

Naturally, it is appealing to minimise this variance jointly for a user-selected choice of values for x, reflected in a distribution $H(x)$. So the goal is to minimise the objective function

$$Q(\xi, \lambda) = \int f^T(x, \lambda) I^{-1}(\xi, \lambda) f(x, \lambda) \, dH(x). \tag{10}$$

To achieve robustness against misspecification of the nonlinear model parameters, we seek Bayesian Q-optimal designs with respect to a prior π, which minimise

$$\Phi_{Q,\pi}(\xi) = \int_\Lambda Q(\xi, \lambda) \, d\pi(\lambda). \tag{11}$$

Similarly, a minimax Q-optimal design minimises

$$\Psi_{Q,\Lambda}(\xi) = \max_{\lambda \in \Lambda} Q(\xi, \lambda). \tag{12}$$

Theorem 3, which is proven in Appendix A.4, establishes the main result of this section, i.e. that the product design of the Bayesian (minimax) Q-optimal designs in the marginal models (4) is Bayesian (minimax) Q-optimal for the additive model (1) in the class of all product designs.

Theorem 3 *Let π be a product prior on $\lambda \in \Lambda$ with marginals π_k on Λ_k, $k = 1, \ldots, K$, and the weighting measure $H(x)$ be a product measure with marginals $H_1(x_1), \ldots, H_K(x_K)$.*

(a) *The product of the Bayesian Q-optimal designs for the single variable models with respect to H_k and π_k, $k = 1, \ldots, K$, is Bayesian Q-optimal within the class of all product designs with respect to H and π.*

(b) *For compact Λ, the product of the minimax Q-optimal designs for the single variable models with respect to H_k and Λ_k, $k = 1, \ldots, K$, is minimax Q-optimal within the class of all product designs with respect to H and Λ.*

5 Discussion

We have illustrated the benefit of using optimal designs through an application to our motivating example on engine mapping. Through our theoretical results, the computational burden to find optimal designs has been reduced considerably, so it is more likely that they will be adopted in industry and generate impact. Even if the complete product designs are too large to be run in practice, it is vital to have a benchmark to compare candidate designs against, in order to avoid inefficient designs being run.

We note that for some applications interactions between the explanatory variables may be present. For linear models with complete product-type interactions, it is well known that for many popular optimality criteria the product of the optimal designs in the marginal models is indeed optimal in the multivariable model. These results, however, do not carry over to partially nonlinear models, for which the complete product-type interaction model constructed from partially nonlinear single variable models is not in general partially nonlinear, so optimal designs for the multivariable model will depend on (some of the) linear parameters, i.e. the coefficients of additive terms in the model.

A Proofs

For clarity of presentation, in what follows we restrict ourselves to proving our results for $K = 2$. The general case $K > 2$ follows by defining meta-variables consisting of more than one single variable and applying the result for $K = 2$.

A.1 Proof of Theorem 1

(a) Let ξ_1, ξ_2 denote the marginals of the design ξ. The special form of the information matrices permits application of Lemma 5.1 in [19], so for all $\lambda = (\lambda_1^T, \lambda_2^T)^T \in \Lambda$

$$|I(\xi, \lambda)| \leq |I_1(\xi_1, \lambda_1)| \, |I_2(\xi_2, \lambda_2)| = |I(\xi_1 \otimes \xi_2, \lambda)|. \tag{13}$$

Using inequality (13) and the assumption that π is a product prior, the following holds:

$$
\begin{aligned}
\int_\Lambda \log |I(\xi_{D,\pi_1}^* \otimes \xi_{D,\pi_2}^*, \lambda)| \, d\pi(\lambda) &\leq \max_\xi \int_\Lambda \log |I(\xi, \lambda)| \, d\pi(\lambda) \\
&\leq \max_{\xi_1, \xi_2} \int_\Lambda \log(|I_1(\xi_1, \lambda_1)| \, |I_2(\xi_2, \lambda_2)|) \, d\pi(\lambda) \\
&= \max_{\xi_1} \int_{\Lambda_1} \log |I_1(\xi_1, \lambda_1)| \, d\pi_1(\lambda_1) \\
&\quad + \max_{\xi_2} \int_{\Lambda_2} \log |I_2(\xi_2, \lambda_2)| \, d\pi_2(\lambda_2) \tag{14} \\
&= \int_{\Lambda_1} \log |I_1(\xi_{D,\pi_1}^*, \lambda_1)| \, d\pi_1(\lambda_1) + \int_{\Lambda_2} \log |I_2(\xi_{D,\pi_2}^*, \lambda_2)| \, d\pi_2(\lambda_2).
\end{aligned}
$$

From the equality in (13) we obtain immediately that

$$\int_\Lambda \log |I(\xi_{D,\pi_1}^* \otimes \xi_{D,\pi_2}^*, \lambda)| \, d\pi(\lambda) = \int_{\Lambda_1} \log |I_1(\xi_{D,\pi_1}^*, \lambda_1)| \, d\pi_1(\lambda_1) + \int_{\Lambda_2} \log |I_2(\xi_{D,\pi_2}^*, \lambda_2)| \, d\pi_2(\lambda_2),$$

so all inequalities in (14) turn into equalities and ξ_π^* is optimal.

(b) Since we consider compact sets Λ_k, $k = 1, 2$, the product set Λ is also compact, and the infimum in (7) is a minimum. Applying (13) and the result for locally D-optimal designs from part (a) of this Theorem, we obtain that for all designs ξ with marginals ξ_1 and ξ_2 and all $\lambda \in \Lambda$:

$$\Phi(\xi, \lambda) \leq \Phi(\xi_1, \lambda_1)\Phi(\xi_2, \lambda_2) = \Phi(\xi_1 \otimes \xi_2, \lambda), \tag{15}$$

where $\Phi(\xi, \lambda) = |I(\xi, \lambda)|/|I(\xi_{D,\lambda}^*, \lambda)|$. Taking the minimum with respect to $\lambda \in \Lambda$ does not change the (in)equalities in (15). Moreover, since Λ is a product set, the two-dimensional minimisation problem can be split up into one-dimensional problems as follows:

$$\min_{\lambda \in \Lambda} \Phi(\xi^*, \lambda) \leq \min_{\lambda_1 \in \Lambda_1} \Phi(\xi_1^*, \lambda_1) \min_{\lambda_2 \in \Lambda_2} \Phi(\xi_2^*, \lambda_2), \tag{16}$$

where ξ^* is standardised maximin D-optimal with respect to Λ with marginals ξ_1^* and ξ_2^*, and

$$\min_{\lambda \in \Lambda} \Phi(\xi_{D,\Lambda_1}^* \otimes \xi_{D,\Lambda_2}^*, \lambda) = \min_{\lambda_1 \in \Lambda_1} \Phi(\xi_{D,\Lambda_1}^*, \lambda_1) \min_{\lambda_2 \in \Lambda_2} \Phi(\xi_{D,\Lambda_2}^*, \lambda_2). \tag{17}$$

From (16), using the optimality of ξ^* in the multivariable model and of ξ_{D,Λ_1}^* and ξ_{D,Λ_2}^* in the single variable models, we find that

$$
\begin{aligned}
\min_{\lambda \in \Lambda} \Phi(\xi_{D,\Lambda_1}^* \otimes \xi_{D,\Lambda_2}^*, \lambda) &\leq \min_{\lambda \in \Lambda} \Phi(\xi^*, \lambda) \leq \min_{\lambda_1 \in \Lambda_1} \Phi(\xi_1^*, \lambda_1) \min_{\lambda_2 \in \Lambda_2} \Phi(\xi_2^*, \lambda_2) \\
&\leq \min_{\lambda_1 \in \Lambda_1} \Phi(\xi_{D,\Lambda_1}^*, \lambda_1) \min_{\lambda_2 \in \Lambda_2} \Phi(\xi_{D,\Lambda_2}^*, \lambda_2).
\end{aligned}
$$

Using (17), all inequalities turn into equalities, which completes the proof of Theorem 1. $\qquad \square$

A.2 Proof of Lemma 1

If $\tilde{C}_{\theta,21} = 0_{(p-s)\times s}$, the inverse of $\tilde{C}_\theta$ also has $0_{(p-s)\times s}$ as its lower left block, and the non-singular $s \times s$-matrix $\tilde{C}_{\theta,11}^{-1}$ as its upper left block. Hence $A^T(\tilde{C}_\theta^{-1})^T = ((\tilde{C}_{\theta,11}^{-1})^T \mid 0_{s\times(p-s)})$, and so

$$|M_{11}^{-1}(\xi,\theta,\lambda)| = |A^T M^-(\xi,\theta,\lambda)A| = |A^T(\tilde{C}_\theta^{-1})^T I^-(\xi,\lambda)\tilde{C}_\theta^{-1}A| = |\tilde{C}_{\theta,11}^{-1}|^2|I_{11}^{-1}(\xi,\lambda)|,$$

where $M_{11}^{-1}(\xi,\theta,\lambda)$ and $I_{11}^{-1}(\xi,\lambda)$ denote the upper left blocks of size $s \times s$ of the matrices $M^-(\xi,\theta,\lambda)$ and $I^-(\xi,\lambda)$, respectively. The assertion of Lemma 1 follows with $c_\theta = |\tilde{C}_{\theta,11}^{-1}|^2$. □

A.3 Proof of Theorem 2

From Lemma 1, we can restrict ourselves to the D_s-criterion for the re-ordered information matrix $I(\xi,\lambda)$, which is at the same time the information matrix for the linear model $\nu(x) = f^T(x,\lambda)\beta$ with iid normal errors and fixed values for λ. (Here $f(x,\lambda)$ is the re-ordered version of the model vector.) From the proof of Theorem 5.13 in [19], we then obtain the inequality

$$\psi_s(I(\xi,\lambda)) \le \psi_s(I_1(\xi_1,\lambda_1))\ \psi_s(I_2(\xi_2,\lambda_2)) = \psi_s(I(\xi_1 \otimes \xi_2,\lambda))$$

for estimating any subset $\varphi = (\varphi_1^T, \varphi_2^T)^T$ of the model parameters not containing the intercept, where ξ_1 and ξ_2 are the marginals of the design ξ, and $\lambda = (\lambda_1^T, \lambda_2^T)^T$. The rest of the proof now follows exactly along the same lines as the proof of Theorem 1 and is therefore omitted. □

A.4 Proof of Theorem 3

Since $g^T(x,\theta,\lambda)M^{-1}(\xi,\theta,\lambda)g(x,\theta,\lambda) = f^T(x,\lambda)I^{-1}(\xi,\lambda)f(x,\lambda)$ a Q-optimal design for model (1) is at the same time Q-optimal for the linear model $\nu(x) = f^T(x,\lambda)\beta$ with iid normal errors and fixed λ. Each single variable model can be expressed by $\nu_k(x_k) = f_k^T(x_k,\lambda_k)\beta_k$ with $f_k^T(x_k) = (1, \tilde{f}_k^T(x_k,\lambda_k))$, so $f^T(x,\lambda) = (1, \tilde{f}_1^T(x_1,\lambda_1), \tilde{f}_2^T(x_2,\lambda_2))$. Lemma 5.5 (ii) in [19] states the form of the covariance matrix $C(\xi,\lambda)$ of the parameter estimators for product designs $\xi = \xi_1 \otimes \xi_2$ in such a model,

$$C(\xi,\lambda) = \left(\begin{array}{c|c|c} C_0(\xi_1 \otimes \xi_2) & -(\int \tilde{f}_1 d\xi_1)^T C_1(\xi_1) & -(\int \tilde{f}_2 d\xi_2)^T C_2(\xi_2) \\ \hline -C_1(\xi_1)(\int \tilde{f}_1 d\xi_1) & C_1(\xi_1) & 0_{p_1 \times p_2} \\ \hline -C_2(\xi_2)(\int \tilde{f}_2 d\xi_2) & 0_{p_2 \times p_1} & C_2(\xi_2) \end{array}\right),$$

where p_k, $k = 1,2$, is the number of parameters in model $\tilde{\nu}_k(x_k) = \tilde{f}_k^T(x_k,\lambda_k)\tilde{\beta}_k$. The covariance matrix in the kth single variable model, $k = 1,2$, is given by

$$C_k(\xi_k,\lambda_k) = \left(\begin{array}{c|c} C_{k,0}(\xi_k) & -(\int \tilde{f}_k d\xi_k)^T C_k(\xi_k) \\ \hline -C_k(\xi_k)(\int \tilde{f}_k d\xi_k) & C_k(\xi_k) \end{array}\right),$$

where $C_0(\xi_1 \otimes \xi_2) = C_{1,0}(\xi_1) + C_{2,0}(\xi_2) - 1$. If $H = H_1 \otimes H_2$ is a product distribution with marginals H_1 and H_2, the objective function for the Q-criterion in the additive model splits according to

$$
\begin{aligned}
\int f^T(x, \lambda) C(\xi, \lambda) f(x, \lambda) dH(x) = {} & H_2(\chi_2) \int f_1^T(x_1, \lambda_1) C_1(\xi_1, \lambda_1) f_1(x_1, \lambda_1) dH_1(x_1) \\
& + H_1(\chi_1) \int f_2^T(x_2, \lambda_2) C_2(\xi_2, \lambda_2) f_2(x_2, \lambda_2) dH_2(x_2) - H(\chi).
\end{aligned}
$$

From this representation, it is obvious that for product designs the local Q-objective function for ξ in the additive model with respect to H and λ is minimised by the product of the Q-optimal designs for the single variable models with respect to H_k and λ_k. Interchanging the integration with respect to $\pi(\lambda)$ (maximisation with respect to $\lambda \in \Lambda$) and the summation of the Q-objective functions in the marginal models yields the desired result for Bayesian (minimax) Q-optimality. $\qquad \square$

References

[1] M. Becka, H.M. Bolt, and W. Urfer. Statistical evaluation of toxicokinetic data. *Environmetrics*, 4:311–322, 1993.

[2] M. Becka and W. Urfer. Statistical aspects of inhalation toxicokinetics. *Environ. Ecol. Stat.*, 3:51–64, 1996.

[3] A. Buja, T. Hastie, and Tibshirani R. Linear smoothers and additive models. *Ann. Statist.*, 17:453–555, 1989.

[4] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

[5] V. C. P. Chen, D. Ruppert, and C. A. Shoemaker. Applying experimental design and regression splines to high dimensional continues state stochastic dynamic programming. *Operations Research*, 47:38–53, 1999.

[6] H. Chernoff. Locally optimal designs for estimating parameters. *Ann. Math. Statist.*, 24:586–602, 1953.

[7] H. Dette. Designing experiments with respect to standardized optimality criteria. *J. Roy. Statist. Soc. Ser. B*, 59:97–110, 1997.

[8] M. L. Dudzinski and R. Mykytowycz. The eye lens as an indicator of age in the wild rabbit in australia. *CSIRO Wildlife Research*, 6:156–159, 1961.

[9] R. L. Eubank. Nonparametric regression and spline smoothing. In *2nd. ed. Statistics: Textbooks and Monographs*, volume 157. Marcel Dekker, New York, 1999.

[10] K.-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments.* Chapman and Hall, London, 2006.

[11] U. Graßhoff, H. Großmann, H. Holling, and R. Schwabe. Design optimality in multi-factor generalized linear models in the presence of an unrestricted quantitative factor. *Journal of Statistical Planning and Inference*, 137:3882–3893, 2007.

[12] D. M. Grove, D. C. Woods, and S. M. Lewis. Multifactor B-spline mixed models in designed experiments for the engine mapping problem. *Journal of Quality Technology*, 36:380–391, 2004.

[13] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

[14] P. D. H. Hill. *D*-optimal designs for partially nonlinear regression models. *Technometrics*, 22:275–276, 1980.

[15] A. I. Khuri. A note on *D*-optimal designs for partially nonlinear regression models. *Technometrics*, 26:59–61, 1984.

[16] R. Put, Q. S. Xu, D. L. Massart, and Y. Vander Heyden. Multivariate adaptive regression splines (mars) in chromatographic quantitative structureretention relationship studies. *Journal of Chromatography A*, 1055:11–19, 2004.

[17] E. Rafajlowicz and W. Myszka. When product type experimental design is optimal? brief survey and new results. *Metrika*, 39:321–333, 1992.

[18] C. Rodriguez and I. Ortiz. *d*-optimum designs in multi-factor models with heteroscedastic errors. *Journal of Statistical Planning and Inference*, 128:623–631, 2005.

[19] R. Schwabe. *Optimum designs for multi-factor models. Lecture Notes in Statistics.* Springer, New York, N. Y., 1996.

[20] S. Siddappa, D. Günther, J. M. Rosenberger, and V. C. P. Chen. Refined experimental design and regression splines method for network revenue management. *Journal of Revenue and Pricing Management*, 6:188–199, 2007.

[21] D. C. Woods, S. M. Lewis, and J. N. Dewynne. Designing experiments for multi-variable *B*-spline models. *Sankhya*, 65:660–677, 2003.