ELSEVIER

Journal of Statistical Planning and Inference 83 (2000) 395–412

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Robust weights and designs for biased regression models: Least squares and generalized M-estimation

Douglas P. Wiens*

Statistics Centre, Department of Mathematical Sciences, University of Alberta, Edmonton, Alta, Canada T6G 2G1

Received 4 June 1997; accepted 1 April 1999

Abstract

We consider an 'approximately linear' regression model, in which the mean response consists of a linear combination of fixed regressors and an unknown additive contaminant. Only the linear component can be modelled by the experimenter. We assume that the experimenter chooses design points and then estimates the regression parameters by weighted least squares or by generalized M-estimation. For this situation we exhibit designs and weights which minimize scalar-valued functions of the covariance matrix of the regression estimates, subject to a requirement of unbiasedness. We report the results of a simulation study in which these designs/weights result in significant improvements, with respect to both bias and mean squared error, when compared to some common competitors.

MSC: primary 62K05, 62F35; secondary 62J05

Keywords: A-optimality; Bounded influence M-estimation; D-optimality; Least median of squares; Legendre polynomial; Multiple regression; Optimal design; Polynomial regression; Q-optimality; Robustness; Weighted least squares

1. Introduction

In this paper we investigate the interplay between the choice of design points for regression based data analyses, and the weights used in weighted least squares (WLS) or generalized M-estimation. The regression model which we envisage is one for which the ordinary least squares (OLS) estimates are biased. Specifically, suppose that the experimenter is to take observations on a random variable Y obeying the 'approximately

* Tel.: (780) 492-4406; fax: (780) 492-4826.
E-mail address: doug.wiens@ualberta.ca (D.P. Wiens)

0378-3758/00/$ - see front matter
PII: S0378-3758(99)00102-0

linear' model

$$Y(\boldsymbol{x}) = \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta} + f(\boldsymbol{x}) + \varepsilon, \tag{1}$$

for some p-dimensional parameter vector $\boldsymbol{\theta}$ and regressors $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x})$, uncorrelated errors ε with common variance σ^2, and an unknown contaminant $f(\boldsymbol{x})$ representing uncertainty about the exact nature of the regression response. For example, the elements of $\boldsymbol{z}$ could be low-degree monomials in the elements of $\boldsymbol{x}$, with $f(\boldsymbol{x})$ being a multinomial of higher degree. Only $\boldsymbol{\theta}$ is estimated, and then $E[Y|\boldsymbol{x}]$ is estimated by $\hat{Y}(\boldsymbol{x}) = \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\hat{\boldsymbol{\theta}}$, so that $\hat{Y}(\boldsymbol{x})$ can be highly biased both for $\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}$ and for $E[Y|\boldsymbol{x}]$. Protection is sought against this bias, and against errors due to random variation.

Our approach is to first determine designs which, for fixed weights, will minimize the maximum bias as $E[Y|\boldsymbol{x}]$ ranges over an L^2-neighbourhood of linear functions. The weights are then chosen to minimize a specified function of the covariance matrix of $\hat{\boldsymbol{\theta}}$. Other methods are possible; one might for instance:

(i) Minimize a function of the covariance matrix of $\hat{\boldsymbol{\theta}}$, subject to the imposition of a bound on a function of the covariance matrix of the regression estimates in a larger model. See Stigler (1971) for an application of this method to OLS estimation in polynomial models.

(ii) Restrict the class of departures from strict linearity in ways other than (1). For such applications to OLS estimation see Pesotchinsky (1982) for multiple linear regression and Liu and Wiens (1997) for polynomial regression.

(iii) Adopt a minimax mean squared error (MSE) approach, maximizing a measure of magnitude of the MSE matrix over possible departures from linearity, and then minimizing this maximum through the choice of weights and design points. Again for OLS, see Wiens (1992) and references cited therein, in particular Huber (1975). For WLS in heteroscedastic models see Wiens (1998).

We believe that the approach adopted here is the simplest of those mentioned, and leads to the most immediately applicable solutions with a minimum of restrictions on the design spaces and response functions involved. We note that the derivation, for *fixed* designs, of weights which are robust against either incompletely specified response functions or influential data points, has been considered by, among others, Hampel et al. (1986), Fedorov et al. (1993) and Simpson and Chang (1997).

The remainder of this section is devoted to a precise description of our class of possible departures from linearity, and a qualitative description of the results obtained.

To obtain (1) we first suppose that the experimenter is to take n uncorrelated observations on a random variable Y whose mean is thought to vary, in an approximately linear manner, with regressors $\boldsymbol{z}(\boldsymbol{x})$: $E[Y|\boldsymbol{x}] \approx \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}$. The sites $\boldsymbol{x}_i$ are chosen from $\mathscr{S}$, a q-dimensional continuous and bounded design space with volume defined by $\Omega^{-1} := \int_{\mathscr{S}} \mathrm{d}\boldsymbol{x}$. We define the 'true' value of $\boldsymbol{\theta}$ by requiring the linear approximation to be most accurate in the L^2-sense:

$$\boldsymbol{\theta} := \arg\min_{\boldsymbol{t}} \int_{\mathscr{S}} (E[Y|\boldsymbol{x}] - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{t})^2 \,\mathrm{d}\boldsymbol{x}.$$

We then define $f(\boldsymbol{x}) = E[Y|\boldsymbol{x}] - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}$ and $\varepsilon(\boldsymbol{x}) = Y(\boldsymbol{x}) - E[Y|\boldsymbol{x}]$, so that (1) holds. These definitions of $\boldsymbol{\theta}$ and of f together imply that

$$\int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x}) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \boldsymbol{0}. \tag{2}$$

In order that errors due to bias will not swamp those due to variance, we shall also assume a uniform bound

$$\int_{\mathscr{S}} f^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \leqslant \eta^2 < \infty. \tag{3}$$

We note that η^2 need not be known in order for our results to be applied.

To quantify the biases and variance/covariance terms, first let ξ be the design measure, i.e. $\xi = n^{-1}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$, where $\delta_{\boldsymbol{x}}$ denotes pointmass of 1 at $\boldsymbol{x}$. For a nonnegative weighting function $w(\boldsymbol{x})$, define vectors and matrices

$$\boldsymbol{b} = \boldsymbol{b}_{f,w,\xi} = \int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x}) f(\boldsymbol{x}) w(\boldsymbol{x}) \xi(\mathrm{d}\boldsymbol{x}),$$

$$\boldsymbol{B} = \boldsymbol{B}_{w,\xi} = \int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x}) \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x}) w(\boldsymbol{x}) \xi(\mathrm{d}\boldsymbol{x}),$$

$$\boldsymbol{D} = \boldsymbol{D}_{w,\xi} = \int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x}) \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x}) w^2(\boldsymbol{x}) \xi(\mathrm{d}\boldsymbol{x}).$$

Of course any implementable design is discrete, and then these quantities may be expressed as averages over the design points. In a more familiar regression notation they are $\boldsymbol{b} = n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{f}$, $\boldsymbol{B} = n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}$ and $\boldsymbol{D} = n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}^2\boldsymbol{Z}$, where $\boldsymbol{Z}$ is the $n \times p$ model matrix with rows $\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x}_i)$, $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with diagonal elements $w(\boldsymbol{x}_i)$ and $\boldsymbol{f}$ is the $n \times 1$ vector with elements $f(\boldsymbol{x}_i)$. The reason for writing them in terms of integrals will become evident in Section 2, where we broaden the class of allowable designs to include continuous measures. Note also that both the weights and design points are to be chosen by the experimenter. This is in contrast to the situation pertaining with heteroscedastic errors, where $w(\boldsymbol{x})$ might represent a known and therefore fixed efficiency function inversely proportional to the error variances.

The weighted least squares estimate

$$\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} = \arg\min_{\theta} \int_{\mathscr{S}} (Y(\boldsymbol{x}) - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta})^2 w(\boldsymbol{x}) \xi(\mathrm{d}\boldsymbol{x}) = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y},$$

where $\boldsymbol{Y}$ is the vector of observations, has bias vector and covariance matrix

$$E[\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} - \boldsymbol{\theta}] = \boldsymbol{B}^{-1}\boldsymbol{b}, \quad \mathrm{COV}[\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}] = \frac{\sigma^2}{n}\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}.$$

Similarly, for a Mallows-type generalized (or 'Bounded Influence') M-estimate defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{GM}} = \arg\min_{\theta} \int_{\mathscr{S}} \rho\left(\frac{Y(\boldsymbol{x}) - \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}}{\sigma}\right) w(\boldsymbol{x}) \xi(\mathrm{d}\boldsymbol{x}) \tag{4}$$

the asymptotic bias of $\sqrt{n}\,\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$ is $\boldsymbol{B}^{-1}\boldsymbol{b}$ and, with $\psi = \rho'$, the asymptotic covariance matrix is $v\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}$ where $v = \sigma^2 E[\psi^2(\varepsilon/\sigma)]/E[\psi'(\varepsilon/\sigma)]^2$. See Hampel et al.

(1986) and Wiens (1996) for background material and details of the asymptotics, respectively.

We will present solutions to the problem of determining a design ξ and weights w to minimize a scalar-valued function $\mathscr{L}(\cdot)$ of the covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}$, or of the asymptotic covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$, subject to the unbiasedness condition

$$\boldsymbol{b}_{f,w,\xi} = \mathbf{0} \quad \text{for all } f. \tag{5}$$

It turns out that the optimal designs have densities inversely proportional to the optimal weighting functions. An algorithm for determining the optimal weights is given in Section 2. Special cases, examples and strategies for implementing such designs are presented in Section 3. All derivations are in the Appendix.

2. Optimal designs and weights

We adopt the viewpoint of *approximate* design theory, and allow as a design measure ξ any probability measure on $\mathscr{S}$. Admissible designs are then not discrete. It can be shown (see Lemma 1 of Wiens 1992) that if $\sup_f ||\boldsymbol{b}||$ is to be finite then ξ must necessarily be absolutely continuous. (Here and elsewhere we use the Euclidean norm $||\cdot||$.) Let k be the density and set $m = kw$, so that $w(\boldsymbol{x})\xi(\mathrm{d}\boldsymbol{x}) = m(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Without loss of generality we may take the average weight $\int_{\mathscr{S}} w(\boldsymbol{x})\xi(\mathrm{d}\boldsymbol{x})$ to be unity, so that m is a density on $\mathscr{S}$. We make the assumptions:

(A1) For each $\boldsymbol{a} \neq \mathbf{0}$, the set $\{\boldsymbol{x} \mid \boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{a} = 0\}$ has Lebesgue measure 0.

(A2) The loss function $\mathscr{L}(\cdot)$, defined on matrices $\boldsymbol{C} \geqslant \mathbf{0}$ (i.e. $\boldsymbol{C}$ positive semidefinite), is

(i) homogeneous, in the sense that minimization of $\mathscr{L}(\kappa\boldsymbol{C})$ is equivalent to minimization of $\mathscr{L}(\boldsymbol{C})$ for $\kappa > 0$ and $\boldsymbol{C} \geqslant \mathbf{0}$;

(ii) isotonic, in that $\boldsymbol{C}_1 \geqslant \boldsymbol{C}_2 \geqslant \mathbf{0}$ implies that $\mathscr{L}(\boldsymbol{C}_1) \geqslant \mathscr{L}(\boldsymbol{C}_2)$;

(iii) Gateaux differentiable; and

(iv) concave: $\mathscr{L}((1-\varepsilon)\boldsymbol{C}_1 + \varepsilon\boldsymbol{C}_2) \geqslant (1-\varepsilon)\mathscr{L}(\boldsymbol{C}_1) + \varepsilon\mathscr{L}(\boldsymbol{C}_2)$ for $\varepsilon \in (0,1)$ and $\boldsymbol{C}_1, \boldsymbol{C}_2 \geqslant \mathbf{0}$.

Assumption (A1) ensures the nonsingularity of a number of relevant matrices, in particular that of $\boldsymbol{A} := \int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x})\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The class of functions satisfying (A2) includes the matrix means $\mathscr{L}(\boldsymbol{C}) = (p^{-1}\operatorname{tr}\boldsymbol{C}^r)^{1/r}$ for $r \leqslant 1$ (see Pukelsheim (1993, Chapter 6). We shall consider in particular the functions $\mathscr{L}_{\mathrm{Q}}(\boldsymbol{C}) = \operatorname{tr}\boldsymbol{A}\boldsymbol{C}$, $\mathscr{L}_{\mathrm{D}}(\boldsymbol{C}) = \log|\boldsymbol{C}|$ and $\mathscr{L}_{\mathrm{A}}(\boldsymbol{C}) = \operatorname{tr}\boldsymbol{C}$. These correspond to the classical Q-, D- and A-optimality problems, and so we adopt the same nomenclature. (We note that Q-optimality is also termed I-optimality by Studden (1977)). The use of these loss functions can be motivated as follows. The mean squared error matrix of $\hat{\boldsymbol{\theta}}$ is

$$\mathrm{MSE} = \mathrm{COV}[\hat{\boldsymbol{\theta}}] + \boldsymbol{B}^{-1}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}^{-1}, \tag{6}$$

and then the integrated mean squared error of $\hat{Y}(\boldsymbol{x})$ as an estimate of $E[Y|\boldsymbol{x}]$ is

$$\text{IMSE} = \int_{\mathscr{S}} E[(\hat{Y}(\boldsymbol{x}) - E[Y|\boldsymbol{x}])^2]\,\mathrm{d}\boldsymbol{x} = \text{tr}(\boldsymbol{A}\cdot\text{MSE}) + \int_{\mathscr{S}} f^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$= \text{tr}(\boldsymbol{A}\cdot\text{COV}[\hat{\boldsymbol{\theta}}]) + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{b} + \int_{\mathscr{S}} f^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

The determinant and trace of the mean squared error matrix of $\hat{\boldsymbol{\theta}}$ are

$$|\text{MSE}| = |\text{COV}[\hat{\boldsymbol{\theta}}]|(1 + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}^{-1}(\text{COV}[\hat{\boldsymbol{\theta}}])^{-1}\boldsymbol{B}^{-1}\boldsymbol{b})$$

and

$$\text{tr}(\text{MSE}) = \text{tr}(\text{COV}[\hat{\boldsymbol{\theta}}]) + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}^{-2}\boldsymbol{b},$$

respectively. Thus minimization of $\mathscr{L}_{\mathrm{Q}}(\boldsymbol{C})$, $\mathscr{L}_{\mathrm{D}}(\boldsymbol{C})$ and $\mathscr{L}_{\mathrm{A}}(\boldsymbol{C})$, where $\boldsymbol{C}$ is a positive scalar multiple of $\text{COV}[\hat{\boldsymbol{\theta}}]$, correspond to minimization of IMSE, |MSE| and tr(MSE) respectively, subject to (5).

Assumption (A2) (iv) excludes the matrix means with $r > 1$, and in particular $(r = \infty)$ the maximum eigenvalue of $\boldsymbol{C}$. See Zhigljavsky (1988) for some relevant techniques for these cases.

Under (A1), condition (5) requires m to be uniform on $\mathscr{S}$.

Theorem 2.1. *In order that the unbiasedness condition* (5) *hold, it is necessary and sufficient that the product* $m(\boldsymbol{x}) = k(\boldsymbol{x})w(\boldsymbol{x})$ *of the design density and weighting function be constant; in order that* $m(\cdot)$ *be a density on* $\mathscr{S}$ *we take*

$$m(\boldsymbol{x}) \equiv \left(\int_{\mathscr{S}} \mathrm{d}\boldsymbol{x}\right)^{-1} = \Omega. \tag{7}$$

Now adopt (7); then $\boldsymbol{B} = \Omega\boldsymbol{A}$ and $\boldsymbol{D} = \Omega\int_{\mathscr{S}} \boldsymbol{z}(\boldsymbol{x})\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})w(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. With

$$\boldsymbol{u}(\boldsymbol{x}) := \boldsymbol{A}^{-1}\boldsymbol{z}(\boldsymbol{x}), \qquad \boldsymbol{C} := \int_{\mathscr{S}} \boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})w(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{8}$$

the asymptotic covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}_{\mathrm{GM}}$ is $(v/\Omega)\boldsymbol{C}$. When $\psi(r) \propto r$ this gives the exact covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}$. Thus, in either case, the mathematical problem is:

Determine nonnegative weights $w(\boldsymbol{x})$ *on* $\mathscr{S}$ *so as to minimize* $\mathscr{L}(\boldsymbol{C})$, *subject to the condition*

$$\int_{\mathscr{S}} w(\boldsymbol{x})^{-1}\,\mathrm{d}\boldsymbol{x} = \Omega^{-1}. \tag{9}$$

Then $k(\boldsymbol{x}) = \Omega w(\boldsymbol{x})^{-1}$ *is the optimal design density.*

Note that the continuous uniform design, employed with OLS, satisfies (7). This design is typically implemented through equally spaced design points. Although the unbiasedness of this continuous design, and the strict optimality of the continuous optimal designs, is lost when such designs are discretized so as to make them implementable, the discretization methods proposed in Section 3.5 lead to design measures which tend weakly to the continuous measures as $n \to \infty$.

Denote by $\mathscr{W}$ the class of nonnegative functions $w(\boldsymbol{x})$ which satisfy (9), hence are positive a.e. $\boldsymbol{x}\in\mathscr{S}$. Let $\mathscr{C}$ be the class of matrices of the form (8) for some $w\in\mathscr{W}$. Let $\mathscr{M}$ be the closed set of positive semidefinite $p\times p$ matrices, $\mathscr{M}^+$ the subset of positive-definite matrices. Then (using (A1)) $\mathscr{C}\subset\mathscr{M}^+\subset\mathscr{M}$. Denote by $w_{\boldsymbol{C}}(\boldsymbol{x})$ any $w\in\mathscr{W}$ which gives $\boldsymbol{C}$ in (8). We are to determine $\boldsymbol{C}\in\mathscr{C}$ and $w_{\boldsymbol{C}}\in\mathscr{W}$ so as to minimize $\mathscr{L}(\boldsymbol{C})$. The solutions to this problem are based on the following result, which gives an algorithm by which any nonoptimal weights may be improved upon.

Theorem 2.2. *For $\boldsymbol{L}_{p\times p}\in\mathscr{M}^+$ define weights $w(\cdot;\boldsymbol{L})\in\mathscr{W}$ by*

$$w(\boldsymbol{x};\boldsymbol{L})=\begin{cases}\dfrac{\Omega\int_{\mathscr{S}}\sqrt{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}}{\sqrt{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})}}, & \text{if } \boldsymbol{u}(\boldsymbol{x})\neq\boldsymbol{0},\\ 0, & \text{otherwise.}\end{cases}$$

For $\boldsymbol{C}\in\mathscr{M}^+$ define $L(\boldsymbol{C})=\mathscr{L}'(\boldsymbol{C})$ by $(L(\boldsymbol{C}))_{ij}=\partial\mathscr{L}(\boldsymbol{C})/\partial c_{ij}$, $1\leqslant i,j\leqslant p$ and define $F:\mathscr{M}^+\to\mathscr{C}$ by

$$F(\boldsymbol{C})_{p\times p}=\int_{\mathscr{S}}\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})w(\boldsymbol{x};L(\boldsymbol{C}))\,\mathrm{d}\boldsymbol{x}. \tag{10}$$

Then $\mathscr{L}(F(\boldsymbol{C}))\leqslant\mathscr{L}(\boldsymbol{C})$. The inequality is strict unless $w_{\boldsymbol{C}}(\boldsymbol{x})=w(\boldsymbol{x};L(\boldsymbol{C}))$ a.e. $\boldsymbol{x}\in\mathscr{S}$, implying that $\boldsymbol{C}=F(\boldsymbol{C})$.

By Theorem 2.2 an optimal covariance matrix $\boldsymbol{C}_*$ is a fixed point of $F(\cdot)$, and optimal weights are $w(\boldsymbol{x};\boldsymbol{L}(\boldsymbol{C}_*))$. In the examples of Section 3 we obtain $\boldsymbol{C}_*$ through the iteration scheme

$$\begin{aligned}&\boldsymbol{C}_{k+1}=F(\boldsymbol{C}_k),\quad k=0,1,\ldots\\ &w_{\boldsymbol{C}_0}(\boldsymbol{x})\equiv 1,\quad \boldsymbol{C}_0=\boldsymbol{A}^{-1}.\end{aligned} \tag{11}$$

Sufficient conditions for the convergence of (11) to a unique minimiser of $\mathscr{L}(\boldsymbol{C})$, from any positive-definite starting point, are that F be a contraction on $\mathscr{C}$:

$$||F(\boldsymbol{C})-F(\boldsymbol{D})||\leqslant\tau||\boldsymbol{C}-\boldsymbol{D}|| \quad \text{for some } \tau<1 \text{ and all } \boldsymbol{C},\boldsymbol{D}\in\mathscr{C}, \tag{12}$$

and that the minimum eigenvalues $\lambda_{\min}(\cdot)$ of matrices $F(\boldsymbol{C})$ be bounded away from 0:

$$\lambda_0:=\inf_{\boldsymbol{C}\in\mathscr{C}}\lambda_{\min}(F(\boldsymbol{C}))>0. \tag{13}$$

For, if (12) holds then any fixed point is unique. That there is a fixed point follows from the observation that (12) implies that $\{\boldsymbol{C}_k\}_{k=0}^{\infty}$ is a Cauchy sequence in $\mathscr{M}^+$, hence has a limit in $\mathscr{M}$. By (13) this limit lies in $\mathscr{M}^+$. Being a fixed point it is a value of F; hence it lies in $\mathscr{C}$. That the fixed point is a minimiser of $\mathscr{L}$ follows from Theorem 2.2.

Note that the convex hull

$$\tilde{\mathscr{C}}=\{\boldsymbol{C}_\varepsilon=(1-\varepsilon)\boldsymbol{C}_0+\varepsilon\boldsymbol{C}_1\mid\boldsymbol{C}_0,\boldsymbol{C}_1\in\mathscr{C},\ 0\leqslant\varepsilon\leqslant 1\}$$

of $\mathscr{C}$ lies in $\mathscr{M}^+$, so that F is defined on $\tilde{\mathscr{C}}$. By the Mean Value Theorem, a sufficient condition for (12) is that the spectral radius be bounded away from one on $\tilde{\mathscr{C}}$:

$$\lambda_1 := \sup_{\boldsymbol{C}_\varepsilon \in \tilde{\mathscr{C}}} \sqrt{\lambda_{\max}(F'^{\mathrm{T}}(\boldsymbol{C}_\varepsilon)F'(\boldsymbol{C}_\varepsilon))} \leqslant \tau < 1. \tag{14}$$

Here F' is the matrix of partial derivatives of F and $\lambda_{\max}(\cdot)$ is the largest eigenvalue. The verification of (13) and (14) is discussed in Section 3 and in Appendix.

3. Special cases and examples

3.1. Q-optimality

For Q-optimality we have $L(\boldsymbol{C}) = \boldsymbol{A}$, so that $F(\boldsymbol{C})$ does not depend on $\boldsymbol{C}$, (13) is immediate, $\tau = 0$ in (13) and the iteration scheme (11) gives the minimizer in one step. (These statements hold whenever $\mathscr{L}(\boldsymbol{C})$ is linear in $\boldsymbol{C}$.) Optimal design densities and weights are

$$k_{\mathrm{Q}}(\boldsymbol{x}) = \frac{\sqrt{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{z}(\boldsymbol{x})}}{\int_{\mathscr{S}} \sqrt{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-1}\boldsymbol{z}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}}, \qquad w_{\mathrm{Q}}(\boldsymbol{x}) = \frac{\Omega}{k_{\mathrm{Q}}(\boldsymbol{x})}. \tag{15}$$

3.2. A-optimality

Similar to the Q-optimality case, for A-optimality we have $L(\boldsymbol{C}) = \boldsymbol{I}$ and so the optimal designs and weights are

$$k_{\mathrm{A}}(\boldsymbol{x}) = \frac{\sqrt{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-2}\boldsymbol{z}(\boldsymbol{x})}}{\int_{\mathscr{S}} \sqrt{\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{A}^{-2}\boldsymbol{z}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}}, \qquad w_{\mathrm{A}}(\boldsymbol{x}) = \frac{\Omega}{k_{\mathrm{A}}(\boldsymbol{x})}. \tag{16}$$

3.3. D-optimality

If $p = 1$, as for straight line regression through the origin, then $w(\boldsymbol{x}; \boldsymbol{L}) \propto 1/|z(\boldsymbol{x})|$ independently of $\boldsymbol{L}$, hence (11) again gives the minimizer in one step. In this case the Q-, A- and D-optimal design densities are all proportional to $|z(\boldsymbol{x})|$. In fact all isotonic criteria coincide if $p = 1$.

For general p Theorem 3.1 establishes (13) and (14) for $\mathscr{L}_{\mathrm{D}}$ and multiple linear regression over a spherical design space, with no interactions between the regressors. For this model $\boldsymbol{z}(\boldsymbol{x}) = (1, \boldsymbol{x}_{q\times 1}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{A}_{q+1\times q+1} = \Omega^{-1} \cdot \mathrm{diag}(1, (q+2)^{-1}, \ldots, (q+2)^{-1})$, and then from (15) and (16) we see that the Q- and A-optimal weights and designs are spherically symmetric. It is thus tempting to conjecture sphericity of the D-optimal designs. A proof of this conjecture has eluded us. However, if $\boldsymbol{T}$ is any orthogonal matrix then the loss $\mathscr{L}_{\mathrm{D}}$ is the same using weights $w(\boldsymbol{T}\boldsymbol{x})$ as using weights $w(\boldsymbol{x})$. There being no *a priori* reason to prefer one direction over another, we *restrict* to weights for which $w(\boldsymbol{T}\boldsymbol{x}) = w(\boldsymbol{x})$ for all orthogonal $\boldsymbol{T}$, i.e. spherically symmetric weights.

Theorem 3.1. *Take* $\mathscr{L} = \mathscr{L}_D$, *so that* $L(\boldsymbol{C}) = \boldsymbol{C}^{-1}$. *For the multiple linear regression model with* $\boldsymbol{z}(\boldsymbol{x}) = (1, \boldsymbol{x}_{q\times 1}^T)^T$ *and* $\mathscr{S} = \{\boldsymbol{x} \mid ||\boldsymbol{x}|| \leqslant 1\}$, *make the restriction to weights* $w(\boldsymbol{x})$ *which are spherically symmetric functions of* $\boldsymbol{x}$. *Then* (13) *and* (14) *hold, so that* (11) *converges to the unique minimiser of* $\mathscr{L}_D(\boldsymbol{C})$ *in* $\mathscr{C}$. *The optimal design has density* $k_D(\boldsymbol{x}) = c_q(1+\gamma_q^2||\boldsymbol{x}||^2)^{1/2}$, *and optimal weights are* $w_D(\boldsymbol{x}) = \Omega k_D(\boldsymbol{x})^{-1}$, *where* $\Omega = \pi^{-q/2}\Gamma(q/2+1)$, $c_q = \Omega/\int_0^1 qu^{q-1}(1+\gamma_q^2u^2)^{1/2}\,\mathrm{d}u$ *and* γ_q *is determined from*

$$\int_0^1 qu^{q-1}\left(\frac{q-\gamma_q^2u^2}{\sqrt{1+\gamma_q^2u^2}}\right)\mathrm{d}u = 0. \tag{17}$$

3.4. Inferences following WLS under homoscedasticity

If WLS is used for reasons other than to address heteroscedasticity, then degrees of freedom are lost in estimating the error variance. This is summarized in the following result.

Theorem 3.2. *Suppose that the data obey the linear model* $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, *where* $\boldsymbol{Z}_{n\times p}$ *has rank* p *and the elements of* $\boldsymbol{\epsilon}$ *are uncorrelated random errors with mean* 0 *and variance* σ^2. *The Weighted Least Squares Estimate* $\hat{\boldsymbol{\theta}}_{WLS} = (\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y}$ *has mean* $\boldsymbol{\theta}$ *and covariance matrix* $(\sigma^2/n)\boldsymbol{C}_0$, *where* $\boldsymbol{C}_0 = n(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z})^{-1}(\boldsymbol{Z}^T\boldsymbol{W}^2\boldsymbol{Z})\times(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z})^{-1}$. *Let* $\boldsymbol{P}_{\boldsymbol{V}}$ *be the projector onto the column space of* $\boldsymbol{V} := (\boldsymbol{Z} \vdots \boldsymbol{W}\boldsymbol{Z})$. *Then an unbiased estimate of* σ^2 *is* $S^2 = ||(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{V}})\boldsymbol{Y}||^2/(n-rk(\boldsymbol{P}_{\boldsymbol{V}}))$. *The vector* $(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{V}})\boldsymbol{Y}$ *is uncorrelated with* $\hat{\boldsymbol{\theta}}_{WLS}$. *If the errors are normally distributed then* $S^2 \sim \sigma^2\chi^2_{n-rk(\boldsymbol{P}_{\boldsymbol{V}})}$, *independently of* $\hat{\boldsymbol{\theta}}_{WLS} \sim \boldsymbol{N}(\boldsymbol{\theta}, (\sigma^2/n)\boldsymbol{C}_0)$.

The projector $\boldsymbol{P}_{\boldsymbol{V}}$ will typically have rank $2p$ when the weights are nonconstant, so that p degrees of freedom are lost in the estimate of σ^2, relative to OLS. Under model (1), S^2 will be positively biased, with bias $E[S^2-\sigma^2] = ||(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{V}})\boldsymbol{f}||^2/(n-rk(\boldsymbol{P}_{\boldsymbol{V}}))$. This is evaluated numerically in the examples below.

Note that S^2 is easily obtained as the mean square of the residuals in a regression of $\boldsymbol{Y}$ on the columns of $\boldsymbol{V}$. Inferences on linear functions of $\boldsymbol{\theta}$ are then carried out in the usual way, with the required change in the degrees of freedom. In particular the t-ratios, computed in Examples 3.1 and 3.2 to estimate by simulation the power of the corresponding single-parameter hypothesis tests, are based on Theorem 3.2.

3.5. Examples

Example 3.1 (*Multiple linear regression*)**.** For the multiple linear regression model detailed in Theorem 3.1 above, the Q-, A- and D-optimal designs can all be written in the same form, viz. with densities $k_q(\boldsymbol{x}) = c_q(1+\gamma_q^2||\boldsymbol{x}||^2)^{1/2}$, differing only in their values of γ_q^2 and hence of the normalizing constants c_q. For Q-optimality we have $\gamma_q^2 = q+2$, while for A-optimality $\gamma_q^2 = (q+2)^2$. See Table 1 for some comparative

Table 1
Values of γ_q^2 (c_q) for multiple linear regression designs and weights

q	Optimality		
	Q	A	D
1	3 (0.3623)	9 (0.2654)	3.787 (0.3428)
2	4 (0.1876)	16 (0.1106)	4.628 (0.1789)
3	5 (0.1212)	25 (0.0613)	5.510 (0.1170)
4	6 (0.0917)	36 (0.0413)	6.423 (0.0893)
5	7 (0.0783)	49 (0.0321)	7.358 (0.0767)
6	8 (0.0737)	64 (0.0279)	8.309 (0.0725)

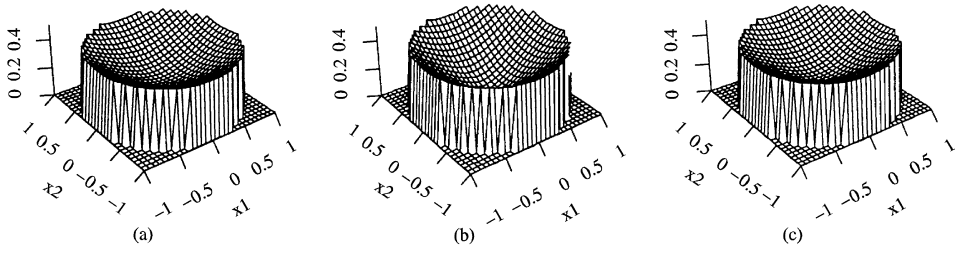

Fig. 1. Optimal design densities for bivariate regression: (a) Q-optimal; (b) A-optimal; (c) D-optimal.

values, and Fig. 1 for plots when $q=2$. From these one sees that the A-optimal designs place relatively more mass near the boundary of $\mathscr{S}$; in this respect they are followed by the D- and then the Q-optimal designs.

To implement these designs, one may use the fact that if $\boldsymbol{X}$ has a spherically symmetric density $k_q(\boldsymbol{x}) = k(||\boldsymbol{x}||)$ on $||\boldsymbol{x}|| \leqslant 1$, then $\boldsymbol{X}/||\boldsymbol{X}||$ is uniformly distributed over the surface of the unit sphere, independently of $U = ||\boldsymbol{X}||$ which has density $g(u) = qu^{q-1}\Omega^{-1}k(u)$ on $[0,1]$. Let $G(\cdot)$ denote the distribution function corresponding to $g(\cdot)$. A possible implementation consists of choosing $a_{n,q}$ design points uniformly distributed over each of the annuli $||\boldsymbol{x}|| = G^{-1}(i/[n/a_{n,q}])$, $i=1,\ldots,[n/a_{n,q}]$, and $n-a_{n,q}[n/a_{n,q}]$ points at $\mathbf{0}$. See Wiens and Zhou (1996) for an example with $a_{n,2}=[\sqrt{n}]$. Note that as long as $a_{n,q}$ and $n/a_{n,q}$ both tend to infinity with n, the discretized measures will tend weakly to the corresponding, optimal, continuous measures.

A plausible alternative is to randomly sample design points from the optimal design measures. Such designs have the same asymptotic properties as the more systematic implementations described above. Simulation studies (details not reported) have however indicated that the sampling variability leads to poor designs, with large values of the loss, with disturbingly high probability. Thus this procedure is not recommended for the well-structured design spaces and response functions considered in these examples. In less structured situations it may be the only feasible approach.

For implementation in an elliptical, uncentered region of interest with generic member $\boldsymbol{v}$, the methods above may be applied to the transformed variables $\boldsymbol{x}_i = \boldsymbol{S}^{-1/2}(\boldsymbol{v}_i - \bar{\boldsymbol{v}})$, where $\boldsymbol{S}$ is the sample covariance matrix computed from the $\boldsymbol{v}_i$. In this case $||\boldsymbol{x}_i||^2$

Table 2
Comparative performance measures for a straight line fit

	Q	A	D	U	M	V
$n=17, a_{n,1}=4$						
Int. MSE	0.240	0.248	0.241	0.258	0.679	1.789
tr(MSE)	0.201	0.201	0.200	0.218	0.398	0.936
$2\sqrt{\lvert\text{MSE}\rvert}$	0.196	0.199	0.196	0.212	0.330	0.467
bias$(\hat{\theta}_0)$[a]	0.134	0.137	0.134	0.160	0.502	0.903
var$(\hat{\theta}_0)$	0.062	0.067	0.063	0.059	0.059	0.059
var$(\hat{\theta}_1)$	0.121	0.116	0.120	0.133	0.088	0.062
bias(S^2)[b]	0.025	0.067	0.031	0.327	0.189	0.138
power[c]	0.465(7)	0.410(7)	0.455(7)	0.534(7)	0.336(7)	0.270(6)
$n=43, a_{n,1}=8$						
Int. MSE	0.089	0.091	0.089	0.094	0.238	0.784
tr(MSE)	0.078	0.078	0.078	0.085	0.143	0.408
$2\sqrt{\lvert\text{MSE}\rvert}$	0.075	0.076	0.075	0.080	0.124	0.191
bias$(\hat{\theta}_0)$[a]	0.054	0.052	0.053	0.071	0.290	0.601
var$(\hat{\theta}_0)$	0.025	0.027	0.025	0.023	0.023	0.023
var$(\hat{\theta}_1)$	0.051	0.049	0.050	0.057	0.036	0.024
bias(S^2)[b]	0.008	0.021	0.010	0.113	0.074	0.021
power[c]	0.536(7)	0.486(7)	0.529(7)	0.558(7)	0.397(7)	0.154(5)

[a] bias$(\hat{\theta}_1)=0$ for all symmetric designs, when $f(x)$ is symmetric.
[b] S^2 is as in Section 3.4 for the weighted designs Q, A and D and is the mean square of the residuals in a regression on the columns of $\boldsymbol{V}=\boldsymbol{Z}$ otherwise.
[c] Standard errors in the third decimal place in parentheses.

is the squared Mahalanobis distance, which can be expressed in terms of the diagonal elements h_{ii} of the 'hat' matrix computed from the regressors $(1,\boldsymbol{v}_i^{\mathrm{T}})$. With $h_{\max}:=\max_{1\leqslant i\leqslant n} h_{ii}$ the optimal weights become

$$w_q(\boldsymbol{v}_i)\propto\left(1+\gamma_q^2\frac{h_{ii}-1/n}{h_{\max}-1/n}\right)^{-1/2}.$$

Table 2 gives some values for the straight line regression model (1) with $\boldsymbol{z}(x)=(1,x)^{\mathrm{T}}$, $|x|\leqslant 1$, and $f(x)=\eta\sqrt{45/8}(x^2-1/3)$ satisfying (2), and (3) with equality. The errors are standard normal. We compare six designs: the Q-, A - and D-optimal designs (Q, A, D) accompanied by the correspondingly optimal weights, the Uniform (U) and Minimax (M) designs without weights and a symmetrization of the variance minimizing design (V), using $n=17$ and $n=43$. Design V places $[n/2]$ design points at each of ± 1, and one at 0. To allow testing for lack of fit, the other designs use $a_{17,1}=4$ and $a_{43,1}=8$, so that for instance when $n=17$ there are 2 points at each of $\pm x_i$, $i=1,\ldots,4$, and one at 0. In this case the x_i are obtained as quartiles of the optimal distributions of $|X|$ as detailed above for Q, A and D; for U they are quartiles of the Uniform distribution on $[0,1]$. Design M is a version of Huber's (1975) minimax design, minimizing the maximum — over all f satisfying (2), and (3) for a particular value of η — integrated mean squared error. We have used the version with density $1.5x^2$; the design points are then the cube roots of those of U.

Table 3
Q-, A- and D-optimal design densities for degree-q polynomial regression

q	$k_Q(x;q)$	$k_A(x;q)$	$k_D(x;q)$
1	$0.362(1+3x^2)^{1/2}$,	$0.265(1+9x^2)^{1/2}$	$0.343(1+3.787x^2)^{1/2}$
2	$0.447(1-2x^2+5x^4)^{1/2}$	$0.140(17-82x^2+125x^4)^{1/2}$	$0.390(1-1.9541x^2+7.540x^4)^{1/2}$
3	$0.130(9+45x^2-165x^4+175x^6)^{1/2}$	$0.022(153+7515x^2-25125x^4+20825x^6)^{1/2}$	$0.111(9+53.094x^2-208.779x^4+272.967x^6)^{1/2}$
4	$0.145(9-36x^2+294x^4-644x^6+441x^8)^{1/2}$	$0.003(40923-651852x^2+3917298x^4-7327852x^6+4234923x^8)^{1/2}$	$0.121(9-35.643x^2+375.113x^4-926.357x^6+731.626x^8)^{1/2}$

We have computed three performance measures from the MSE matrix (6). This matrix is evaluated at the discretized designs described above, so that $\mathrm{COV}[\hat{\boldsymbol{\theta}}]$ is as given in Theorem 3.2, $\boldsymbol{b}=n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{f}$ and $\boldsymbol{B}=n^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}$. The performance measures are Int. MSE $:=\mathrm{tr}(\boldsymbol{A}\cdot\mathrm{MSE})$ (= the IMSE of $\hat{Y}(\boldsymbol{x})$ as an estimate of $\boldsymbol{z}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\theta}$), the trace of MSE and the normalized determinant $2\sqrt{|\mathrm{MSE}|}$; we also present the individual biases and variances. The power of the level 0.05 t-test of H_0: $\theta_2=0$, after fitting a quadratic response, was obtained by simulation. We chose η in such a way that $\sigma^2/n\eta^2=0.15$; this gave powers around 1/2 in most cases. For Q, A and D these powers appear to be closely related to the bias of S^2, as one would expect.

Of the three new designs the best performers are Q and D, although all three gave very similar results with improvements of 5–8% on U with respect to the MSE — based measures; much more with respect to bias. Design V is clearly unable to handle the presence of the contamination $f(x)$ in the response function. As might be expected, the performance of M falls into the middle ground between that of U and that of V.

For the $n=17$ case detailed in Table 2, A is outperformed (very slightly — in the third or fourth decimal places only) by D and Q with respect to the trace of MSE — the measure for which A is ostensibly optimal. This is evidently a result of the discretization of the continuous designs, and is no longer the case at $n=43$ (where the superiority of A with respect to tr(MSE) is again only in the fourth decimal place). The derivation of optimality criteria for the discretization methods is the subject of continuing research.

Example 3.2 (*Polynomial regression*). For $\boldsymbol{z}(x)=(1,x,\ldots,x^q)^{\mathrm{T}}$ (degree-q polynomial regression) with $q=1,2,3$ and 4 the densities $k_Q(x)=k_Q(x;q)$ given by (15), $k_A(x)=k_A(x;q)$ given by (16) and $k_D(x)=k_D(x;q)$ given by Theorem 3.1 are displayed in Table 3.

Given the lack of an analogue of Theorem 3.1 when $q>1$, the D-optimality problem here rests on somewhat intuitive grounds. Note however that the first iteration in (11), using $\mathscr{L}_D$, gives weights $w_{C_1}(\boldsymbol{x})=w_Q(\boldsymbol{x})$. There is little change in subsequent iterations, and we find (see Fig. 2) that the designs optimal for $\mathscr{L}_Q$ are almost identical to those for $\mathscr{L}_D$. This is consistent with the findings of Studden (1977) who noted, for polynomial regression, an analogous relationship between the classical D- and Q-optimal designs ξ_D and ξ_Q, i.e. the discrete measures minimizing the determinant of the covariance

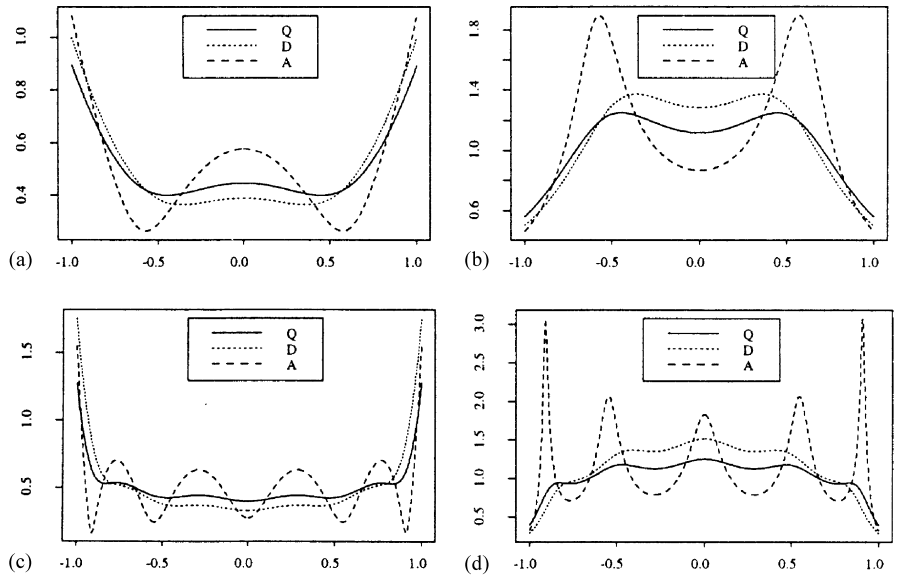

Fig. 2. Optimal designs and weights for degree-q polynomial regression: (a) design densities, $q = 2$; (b) weights, $q = 2$; (c) design densities, $q = 5$; (d) weights, $q = 5$.

matrix of $\hat{\boldsymbol{\theta}}$ and the integrated variance of $\hat{Y}(x)$, respectively. Wiens (1992) made similar findings for minimax robust designs.

In the case of Q-optimality there is an interesting connection to the Legendre polynomials and to these classical designs ξ_D and ξ_Q.

Theorem 3.3. *Denote by* $P_m(x)$ *the* mth *degree Legendre polynomial on* $[-1,1]$, *normalized by* $\int_{-1}^{1} P_m^2(x)\,dx = (m+0.5)^{-1}$. *Define a density on* $[-1,1]$ *by* $h_q(x) = (q+1)^{-1}\boldsymbol{z}^{\mathrm{T}}(x)\boldsymbol{A}^{-1}\boldsymbol{z}(x)$. *Then* $h_q(x) = 0.5(P_q(x)P'_{q+1}(x) - P'_q(x)P_{q+1}(x))$ *and*

$$k_Q(x;q) \propto h_q(x)^{1/2}, \tag{18}$$

$$\lim_{q\to\infty} k_Q(x;q) = \frac{(1-x^2)^{-1/4}}{\sqrt{2}\beta(\frac{3}{4},\frac{3}{4})}. \tag{19}$$

It can be shown that the local maxima of $h_q(x)$, hence those of $k_Q(x;q)$, are the zeros of $(1-x^2)P'_q(x)$. These are precisely the points of support of ξ_D. In this sense $k_Q(x;q)$ is a smoothed version of ξ_D, which has the limiting density $(1-x^2)^{-1/2}/\pi = \lim_{q\to\infty} h_q(x)$. The limiting density of ξ_Q is however the same as the limit (18) of $k_Q(x;q)$ — see Studden (1977). See Fig. 2 for plots in the cases $q = 2$ and $q = 5$.

Table 4 gives some comparative values for a quadratic fit, i.e. (1) with $\boldsymbol{z}(x) = (1,x,x^2)^{\mathrm{T}}$, $|x| \leqslant 1$ and $f(x) = \eta\sqrt{175/8}(x^3 - 0.6x)$. We took $n = 21$. The points comprising designs Q, A, D and U were obtained as quantiles of the densities k_Q, k_A and k_D given above, and of the Uniform density, starting and ending at ± 1, with 3 replicates at each

Table 4
Comparative performance measures for a quadratic fit

	Q	A	D	U	M	V
Int. MSE	0.395	0.402	0.408	0.437	0.489	0.879
tr(MSE)	0.639	0.643	0.657	0.704	0.760	1.262
$3\sqrt[3]{\lvert\text{MSE}\rvert}$	0.418	0.439	0.410	0.482	0.508	0.629
bias($\hat{\theta}_1$)[a]	0.237	0.283	0.199	0.406	0.494	0.913
var($\hat{\theta}_0$)	0.131	0.123	0.146	0.111	0.121	0.143
var($\hat{\theta}_1$)	0.105	0.110	0.104	0.107	0.096	0.071
var($\hat{\theta}_2$)	0.347	0.329	0.367	0.321	0.299	0.214
bias(S^2)[b]	0.000	0.023	0.010	0.257	0.341	0.000
power[c]	0.443(8)	0.371(8)	0.424(8)	0.524(8)	0.646(8)	0.00(0)

[a] $\text{bias}(\hat{\theta}_0) = \text{bias}(\hat{\theta}_2) = 0$ for all symmetric designs, when $f(x)$ is symmetric.
[b] S^2 is as in Section 3.4 for the weighted designs Q, A and D and is the mean square of the residuals in a regression on the columns of $\boldsymbol{V} = \boldsymbol{Z}$ otherwise.
[c] Standard errors in the third decimal place in parentheses.

site. The variance minimizing design V has 7 replicates at each of the sites $-1, 0, 1$. It remains an open problem to derive the design which is minimax against departures, given by (2) and (3), from a qth degree polynomial response with $q > 1$. However, Liu and Wiens (1997) obtained designs which are minimax against departures of the form (1) but with (2) and (3) replaced by the condition $f(x) = x^{q+1}\psi(x)$, for some unknown function $\psi(\cdot)$ assumed to be continuous and bounded above and below (by ± 1, without loss of generality) on $[-1, 1]$. Here we use the version of the resulting design M which is minimax for $q = 3$ and $\sigma^2 = 0.01n$. It has $M(\pm 1) = 0.1703$ and $M(\pm 0.445) = 0.3297$; after rounding and symmetrization this translates into 4 points at each of ± 1, 6 points at each of ± 0.445, and 1 point at 0. The power of the level 0.05 t-test of H_0: $\theta_3 = 0$ was also simulated after fitting a cubic model. This was not possible with V, nor would it have been possible with the design M minimax in the above sense for $q = 2$. We chose η to satisfy $\sigma^2/n\eta^2 = 0.20$. The relative performance of the designs was quite similar to that exhibited in Example 3.1 above. In this case however the superiority of Q over A was at least partially an effect of the replication, in that for *unreplicated* designs with the same values of n, η and σ^2, design A had the smallest value of tr(MSE) and D had the smallest value of $3\sqrt{|\text{MSE}|}$, while Q had the smallest value of Int. MSE. With the unreplicated designs the bias of S^2 increased markedly.

4. Summary

We have obtained regression designs and weights which minimize functions of the covariance matrix of the estimates of the regression parameters, subject to first minimizing the bias, in an approximately linear model. Our results apply to WLS and to GM estimates. Algorithms and explicit solutions have been presented, for

arbitrary response functions and design spaces. Methods of implementation have been discussed. The solutions have been shown to compare very favourably with existing techniques.

Acknowledgements

We are grateful for the very thorough reviews, by two anonymous referees, of a previous version of this paper, and for the helpful comments of these referees and an Associate Editor. This research is supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Derivations

Theorem 2.1 is proved as Theorem (2b) in Wiens (1998). The proof of Theorem 2.2 requires a preliminary result.

Lemma A.1. *For any* $\boldsymbol{L}\geqslant 0$ *define* $\boldsymbol{F}_{\boldsymbol{L}}=\int_{\mathscr{S}}\boldsymbol{u}(x)u^{\mathrm{T}}(\boldsymbol{x})w(\boldsymbol{x};\boldsymbol{L})\,\mathrm{d}\boldsymbol{x}$. *Then*

$$\operatorname{tr}\boldsymbol{L}\boldsymbol{F}_{\boldsymbol{L}}\leqslant\operatorname{tr}\boldsymbol{L}\boldsymbol{C}_{\varepsilon}$$

for any $\boldsymbol{C}_{\varepsilon}\in\tilde{\mathscr{C}}$. *When* $\varepsilon=0$ *the inequality is strict unless* $w(\boldsymbol{x};\boldsymbol{L})=w_{\boldsymbol{C}_0}(\boldsymbol{x})$ *a.e.* $\boldsymbol{x}\in\mathscr{S}$, *implying that* $\boldsymbol{C}_0=\boldsymbol{F}_{\boldsymbol{L}}$.

Proof. Let

$$k_{\varepsilon}(\boldsymbol{x})=\frac{\Omega}{w_{\varepsilon}(\boldsymbol{x})}\Big/\int_{\mathscr{S}}\frac{\Omega}{w_{\varepsilon}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}$$

be the density induced by $w_{\varepsilon}(\boldsymbol{x})=(1-\varepsilon)w_{\boldsymbol{C}_0}(x)+\varepsilon w_{\boldsymbol{C}_1}(x)$. Note that, by virtue of the convexity of the mapping $t\to t^{-1}$, we have that $\int_{\mathscr{S}}(\Omega/w_{\varepsilon}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\leqslant 1$. Then

$$\begin{aligned}
\operatorname{tr}\boldsymbol{L}\boldsymbol{F}_{\boldsymbol{L}} &= \Omega\left(\int_{\mathscr{S}}\sqrt{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\right)^2\\
&= \Omega^{-1}\left(\int_{\mathscr{S}}\sqrt{\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})}w_{\varepsilon}(\boldsymbol{x})k_{\varepsilon}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right)^2\left(\int_{\mathscr{S}}\frac{\Omega}{w_{\varepsilon}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\right)^2\\
&\leqslant \Omega^{-1}\int_{\mathscr{S}}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})w_{\varepsilon}^2(\boldsymbol{x})k_{\varepsilon}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\left(\int_{\mathscr{S}}\frac{\Omega}{w_{\varepsilon}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\right)^2\\
&= \int_{\mathscr{S}}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{L}\boldsymbol{u}(\boldsymbol{x})w_{\varepsilon}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\int_{\mathscr{S}}\frac{\Omega}{w_{\varepsilon}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\\
&\leqslant \operatorname{tr}\boldsymbol{L}\boldsymbol{C}_{\varepsilon}.
\end{aligned}$$

When $\varepsilon=0$ the first inequality above is strict unless $w(\boldsymbol{x};\boldsymbol{L})=w_{\boldsymbol{C}_0}(\boldsymbol{x})$ a.e. $\boldsymbol{x}\in\mathscr{S}$. □

Proof of Theorem 2.2. For $\varepsilon \in [0,1]$ define $g(\varepsilon) = \mathscr{L}(\boldsymbol{C} + \varepsilon(F(\boldsymbol{C}) - \boldsymbol{C}))$. The assumptions imply that g is differentiable, with $g'(0) \geqslant g(1) - g(0)$. Equivalently,

$$\operatorname{tr} L(\boldsymbol{C})(F(\boldsymbol{C}) - \boldsymbol{C}) \geqslant \mathscr{L}(F(\boldsymbol{C})) - \mathscr{L}(\boldsymbol{C})$$

and by Lemma A.1 the left-hand term above is $\leqslant 0$; strictly negative unless $w(\boldsymbol{x}; \boldsymbol{L}) = w_{\boldsymbol{C}}(\boldsymbol{x})$ a.e. $\boldsymbol{x} \in \mathscr{S}$. □

Proof of Theorem 3.1. The restriction to spherically symmetric weights ensures that each member of $\mathscr{C}$ is of the form $\alpha \oplus \beta \boldsymbol{I}_q := \operatorname{diag}(\alpha, \beta, \beta, \ldots, \beta)$. Then $F(\boldsymbol{C}) = \Omega(f_1(\boldsymbol{C}) \oplus f_2(\boldsymbol{C}) \cdot \boldsymbol{I}_q)$, where

$$f_1(\boldsymbol{C}) = \int_{\mathscr{S}} w(\boldsymbol{x}; \boldsymbol{C}^{-1}) \Omega \, \mathrm{d}\boldsymbol{x}, \quad f_2(\boldsymbol{C}) = \frac{(q+2)^2}{q} \int_{\mathscr{S}} ||\boldsymbol{x}||^2 w(\boldsymbol{x}; \boldsymbol{C}^{-1}) \Omega \, \mathrm{d}\boldsymbol{x}. \quad \text{(A.1)}$$

We shall make frequent use of the identity $\int_{\mathscr{S}} h(||\boldsymbol{x}||) \Omega \, \mathrm{d}\boldsymbol{x} = \int_0^1 h(u) q u^{q-1} \, \mathrm{d}u$.

We first establish (13). Denote expectation with respect to the uniform density Ω by E_Ω. Jensen's Inequality yields

$$f_1(\boldsymbol{C}) = E_\Omega[(w(\boldsymbol{x}; \boldsymbol{C}^{-1})^{-1})^{-1}] \geqslant (E_\Omega[w(\boldsymbol{x}; \boldsymbol{C}^{-1})^{-1}])^{-1} = 1,$$

using (9). Similarly, with E_k denoting expectation with respect to the density $\Omega w(\boldsymbol{x}; \boldsymbol{C}^{-1})^{-1}$ we have

$$f_2(\boldsymbol{C}) = \frac{(q+2)^2}{q} E_k[||\boldsymbol{x}||^2 w^2(\boldsymbol{x}; \boldsymbol{C}^{-1})] \geqslant \frac{(q+2)^2}{q} (E_k[||\boldsymbol{x}|| w(\boldsymbol{x}; \boldsymbol{C}^{-1})])^2$$

$$= \frac{(q+2)^2}{q} \Omega^2 \left(\int_{\mathscr{S}} ||\boldsymbol{x}|| \, \mathrm{d}\boldsymbol{x} \right)^2 = \frac{q(q+2)^2}{(q+1)^2} > 1.$$

Thus $\lambda_0 \geqslant \Omega > 0$.

To establish (14) we first note that (A.1) continues to hold with $\boldsymbol{C}$ replaced by $\boldsymbol{C}_\varepsilon = \alpha \oplus \beta \boldsymbol{I}_q \in \widetilde{\mathscr{C}}$. With $\gamma := (q+2)\sqrt{\alpha/\beta}$ it then turns out that the various quantities of interest are expressible in terms of the functions $I_j(\gamma; q) = \int_0^1 (1 + \gamma^2 u^2)^{1/2-j} q u^{q-1} \, \mathrm{d}u$, $j = 0, 1, 2$. We find that $w(\boldsymbol{x}; \boldsymbol{C}_\varepsilon^{-1}) = I_0(\gamma; q)/\sqrt{1 + \gamma^2 ||\boldsymbol{x}||^2}$ and that

$$(f_1(\boldsymbol{C}_\varepsilon),\ f_2(\boldsymbol{C}_\varepsilon))^{\mathrm{T}} =: \boldsymbol{g}(\gamma) = (I_0(\gamma; q) I_1(\gamma; q), \frac{(q+2)^2}{q\gamma^2} I_0(\gamma; q)(I_0(\gamma; q) - I_1(\gamma; q)))^{\mathrm{T}}.$$

We may identify F with the map $(\alpha, \beta) \to \Omega \boldsymbol{g}(\gamma)$, whence $F'(\boldsymbol{C}_\varepsilon) = \Omega \boldsymbol{g}'(\gamma)(\partial\gamma/\partial\alpha, \beta)$. Then $F'^{\mathrm{T}}(\boldsymbol{C}_\varepsilon) F'(\boldsymbol{C}_\varepsilon)$ is of rank one, with maximum eigenvalue

$$\lambda_{\max}(\gamma) = \Omega^2 ||\boldsymbol{g}'(\gamma)||^2 \left|\left| \left(\frac{\partial\gamma}{\partial\alpha, \beta} \right)^{\mathrm{T}} \right|\right|^2. \quad \text{(A.2)}$$

From Lemma A.1 with $\boldsymbol{L} = \boldsymbol{C}_\varepsilon^{-1}$ we have $q + 1 \geqslant \operatorname{tr}(\boldsymbol{C}_\varepsilon^{-1} F(\boldsymbol{C}_\varepsilon)) = (\Omega/\alpha) I_0^2(\gamma)$, so that

$$\left|\left| \left(\frac{\partial\gamma}{\partial\alpha, \beta} \right)^{\mathrm{T}} \right|\right| = \left|\left| \frac{\gamma}{2\alpha} \left(1, \frac{\alpha}{\beta} \right)^{\mathrm{T}} \right|\right| = \frac{\gamma}{2\alpha} \sqrt{1 + \left(\frac{\gamma}{q+2} \right)^4} \leqslant \frac{(q+1)\gamma}{2\Omega I_0^2(\gamma; q)} \sqrt{1 + \left(\frac{\gamma}{q+2} \right)^4}. \quad \text{(A.3)}$$

Using the identities $\gamma I_0'(\gamma;q)=I_0(\gamma;q)-I_1(\gamma;q)$, $\gamma I_1'(\gamma;q)=I_2(\gamma;q)-I_1(\gamma;q)$ we calculate that

$$||\boldsymbol{g}'(\gamma)|| = \frac{(I_0(\gamma;q)I_2(\gamma;q) - I_1^2(\gamma;q))}{\gamma}\sqrt{1+\frac{(q+2)^4}{q^2\gamma^4}}. \tag{A.4}$$

With k_ε as in Lemma A.1 we have

$$\frac{\beta}{\alpha} = \frac{(q+2)^2}{q}\int_{\mathscr{S}} ||\boldsymbol{x}||^2 k_\varepsilon(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \leqslant \frac{(q+2)^2}{q},$$

so that $\gamma \geqslant \sqrt{q}$. Now define

$$\tau_1(\gamma;q) = \frac{q+1}{2(q+2)^2}\sqrt{\left(1+\frac{(q+2)^4}{q^2\gamma^4}\right)\left(1+\frac{(q+2)^4}{\gamma^4}\right)},$$

$$\tau_2(\gamma;q) = \frac{1}{I_0^2(\gamma;q)}, \quad \tau_3(\gamma;q) = \gamma^2(I_0(\gamma;q)I_2(\gamma;q) - I_1^2(\gamma;q)).$$

Combining (A.2)–(A.4) then gives

$$\lambda_1 \leqslant \sup_{\gamma\geqslant\sqrt{q}}\sqrt{\lambda_{\max}(\gamma)} \leqslant \prod_{i=1}^{3}\sup_{\gamma\geqslant\sqrt{q}}\tau_i(\gamma;q).$$

We present the rest of the details for $q \geqslant 4$ only. A separate but similar case-by-case analysis gives the result for $q = 1,2,3$. An inspection of $\tau_i(\gamma;q)$ and differentiation of $\tau_i(\sqrt{q};q)$ reveals that they are decreasing in γ and q respectively, so that

$$\tau_i(\gamma;q) \leqslant \tau_i(\sqrt{q};q) \leqslant \tau_i(2;4), \quad i = 1,2.$$

We write the remaining term as

$$\tau_3(\gamma;q) = E[X]E[X^{-3}] - (E[X^{-1}])^2,$$

where $X = \sqrt{U^2+\gamma^{-2}}$ and U has density qu^{q-1} on $[0,1]$. We then calculate that

$$\frac{\partial\tau_3(\gamma;q)}{\partial\gamma} = \frac{3}{\gamma^3}(\mathrm{cov}[X^{-1},X^{-3}] - \mathrm{cov}[X,X^{-5}]).$$

This is positive since $\mathrm{cov}[X^{-1},X^{-3}] > 0 > \mathrm{cov}[X,X^{-5}]$, so that

$$\tau_3(\gamma;q) \leqslant \tau_3(\infty;q) = \frac{4q^2}{(q-3)(q-1)^2(q+1)} \leqslant \tau_3(\infty;4).$$

Thus

$$\lambda_1 \leqslant \tau_1(2;4)\tau_2(2;4)\tau_3(\infty;4) \approx 0.612,$$

establishing (14). Eq. (17) follows upon equating γ to $(q+2)\sqrt{f_1(\boldsymbol{C})/f_2(\boldsymbol{C})}$ (this equality is implied by the fixed point property) and simplifying. □

Proof of Theorem 3.2. The derivation of the moments of $\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}$ is straightforward. We then note that

$$(\boldsymbol{I}-\boldsymbol{P_V})\boldsymbol{Z} = \boldsymbol{0}_{n\times p}, \quad (\boldsymbol{I}-\boldsymbol{P_V})\boldsymbol{W}\boldsymbol{Z} = \boldsymbol{0}_{n\times p} \tag{A.5}$$

by the definition of $\boldsymbol{P}_V$. Thus $E[(\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{Y}]=\boldsymbol{0}$ and

$$\begin{aligned} E[S^2] &= \frac{\mathrm{tr}\{(\boldsymbol{I}-\boldsymbol{P}_V)E[\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}]\}}{n-rk(\boldsymbol{P}_V)} \\ &= \frac{\mathrm{tr}\{(\boldsymbol{I}-\boldsymbol{P}_V)(\sigma^2\boldsymbol{I}+E[\boldsymbol{Y}]E[\boldsymbol{Y}^{\mathrm{T}}])\}}{n-rk(\boldsymbol{P}_V)} \\ &= \sigma^2, \end{aligned}$$

where the last step uses the first equality in (A.5). Similarly, since $E[(\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{Y}]=\boldsymbol{0}$, we have

$$\begin{aligned} \mathrm{COV}[(\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{Y},\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}] &= E[(\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z})^{-1}] \\ &= (\boldsymbol{I}-\boldsymbol{P}_V)(\sigma^2\boldsymbol{I}+E[\boldsymbol{Y}]E[\boldsymbol{Y}^{\mathrm{T}}])\boldsymbol{W}\boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z})^{-1} \\ &= \boldsymbol{0}_{n\times p}, \end{aligned}$$

using both equalities in (A.5). □

Proof of Theorem 3.3. Define $\boldsymbol{p}(x)=(P_0(x),\ldots,P_q(x))^{\mathrm{T}}$, and let $\boldsymbol{P}_{q+1\times q+1}$ be the matrix of coefficients of the Legendre polynomials, defined through $\boldsymbol{p}(x)=\boldsymbol{P}\boldsymbol{z}(x)$. Then with

$$\boldsymbol{D}=\mathrm{diag}(2,\ldots,(i+0.5)^{-1},\ldots,(q+0.5)^{-1}))$$

we have

$$\boldsymbol{D}=\int_{-1}^{1}\boldsymbol{p}(x)\boldsymbol{p}^{\mathrm{T}}(x)\,\mathrm{d}x=\boldsymbol{P}\int_{-1}^{1}\boldsymbol{z}(x)\boldsymbol{p}^{\mathrm{T}}(x)\,\mathrm{d}x=\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{\mathrm{T}}$$

so that $\boldsymbol{A}=\boldsymbol{P}^{-1}\boldsymbol{D}\boldsymbol{P}^{-1^{\mathrm{T}}}$. We then calculate that

$$h_q(x)=(q+1)^{-1}\sum_{i=0}^{q}(i+0.5)P_i^2(x),$$

and formula 8.915.1 of Gradshteyn and Ryzhik (1980) gives $h_q(x)=0.5(P_q(x)P'_{q+1}(x)-P'_q(x)P_{q+1}(x))$. A standard asymptotic expansion for Legendre polynomials — formula 8.965 of Gradshteyn and Ryzhik (1980) — yields (19). □

References

Fedorov, V.V., Hackl, P., Müller, W.G., 1993. Moving local regression: the weight function. J. Nonparametric Statist. 3, 355–368.

Gradshteyn, I.S., Ryzhik, I.M., 1980. Table of Integrals, Series, and Products. Academic Press, Toronto.

Hampel, F.R., Ronchetti, E., Rousseeuw, R.J., Stahel, W., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, Toronto.

Huber, P.J., 1975. Robustness and designs. In: Srivastava, J.N. (Ed.), A Survey of Statistical Design and Linear Models. North-Holland, Amsterdam, pp. 287–303.

Liu, S.X., Wiens, D.P., 1997. Robust designs for approximately polynomial regression. J. Statist. Plann. Inference 64, 369–381.

Pesotchinsky, L., 1982. Optimal robust designs: linear regression in R^k. Ann. Statist. 10, 511–525.

Pukelsheim, F., 1993. Optimal design of experiments. Wiley, Toronto.

Simpson, D.G., Chang, Y.-C.I., 1997. Reweighting approximate GM estimators: asymptotics and residual-based graphics. J. Statist. Plann. Inference 57, 273–293.

Stigler, S., 1971. Optimal experimental design for polynomial regression. J. Am. Statist. Assoc. 66, 311–318.

Studden, W.J., 1977. Optimal designs for integrated variance in polynomial regression. In: Gupta, S.S., Moore, D.S. (Eds.), Statistical Decision Theory and Related Topics II. Academic Press, New York, pp. 411–420.

Wiens, D.P., 1992. Minimax designs for approximately linear regression. J. Statist. Plann. Inference 31, 353–371.

Wiens, D.P., 1996. Asymptotics of generalized M-estimation of regression and scale with fixed carriers, in an approximately linear model. Statist. Probab. Lett. 30, 271–285.

Wiens, D.P., 1998. Minimax robust designs and weights for approximately specified regression models with heteroscedastic errors. J. Am. Statist. Assoc. 93, 1440–1450.

Wiens, D.P., Zhou, J., 1996. Minimax regression designs for approximately linear models with autocorrelated errors. J. Statist. Plann. Inference 55, 95–106.

Zhigljavsky, A.A., 1988. Optimal designs for estimating several integrals In: Dodge, Y., Fedorov, V.V., Wynn, H.P. (Eds.), Optimal Designs and Analysis of Experiments. North-Holland, Amsterdam, pp. 81–95.