Computational Statistics and Data Analysis 53 (2008) 312–320

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Improving updating rules in multiplicative algorithms for computing *D*-optimal designs

Holger Dette [a,*], Andrey Pepelyshev [b], Anatoly Zhigljavsky [c]

[a] *Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany*

[b] *St. Petersburg State University, Department of Mathematics, St. Petersburg, Russia*

[c] *Cardiff University, School of Mathematics, Cardiff CF24 4AG, UK*

ARTICLE INFO

Article history:
Received 10 March 2008
Accepted 30 September 2008
Available online 9 October 2008

ABSTRACT

A class of multiplicative algorithms for computing *D*-optimal designs for regression models on a finite design space is discussed and a monotonicity result for a sequence of determinants obtained by the iterations is proved. As a consequence the convergence of the sequence of designs to the *D*-optimal design is established. The class of algorithms is indexed by a real parameter and contains two algorithms considered previously as special cases. Numerical results are provided to demonstrate the efficiency of the proposed methods. Finally, several extensions to other optimality criteria are discussed.

1. Introduction

Consider the common linear regression model

$$y = \theta^{\mathrm{T}} u + \varepsilon, \tag{1}$$

where $\theta = (\theta_0, \ldots, \theta_{m-1})^{\mathrm{T}} \in \mathbb{R}^m$ is a vector of unknown parameters, u denotes the vector of explanatory variables and ε is a random error. We assume that $\mathcal{U} = \{u_1, \ldots, u_n\} \subset \mathbb{R}^m$ is a finite design space (a generalization to arbitrary design spaces is straightforward, see Remark 4 in Section 2) and that different observations are uncorrelated; the mean and variance of the errors are 0 and $\sigma^2 > 0$, respectively. Following Kiefer (1974) we call any probability measure ξ on $\mathcal{U}$ a design. If N observations can be taken and the design ξ puts masses $w_1, \ldots, w_n$ at the points $u_1, \ldots, u_n$, then a rounding procedure is applied to obtain integers $n_i \approx w_i N$ with $\sum_{j=1}^n n_i = N$, and the experimenter takes approximately n_i observations at each u_i $(i = 1, \ldots, n)$ [see Pukelsheim and Rieder (1992) for more details and some references]. For a design ξ, the information matrix in the model (1) is defined by

$$M(\xi) = \sum_{i=1}^{n} w_i u_i u_i^{\mathrm{T}},$$

and its inverse is approximately proportional to the covariance matrix of the least squares estimate for the parameter θ. An optimal design maximizes an appropriate function of the information matrix [see e.g. Silvey (1980) or Pukelsheim (1993)]. Numerous optimality criteria have been proposed in the literature to discriminate between competing designs. In the present paper we mainly consider the *D*-optimality criterion, which determines the design ξ^* such that the determinant $|M(\xi)|$ is maximal. Such a design minimizes the volume of the ellipsoid of concentration for the vector θ of unknown

* Corresponding author.
E-mail addresses: holger.dette@rub.de (H. Dette), andrey@ap7236.spb.edu (A. Pepelyshev), ZhigljavskyAA@cf.ac.uk (A. Zhigljavsky).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.10.002

parameters. In most cases of practical interest, D-optimal designs have to be determined numerically and several algorithms have been proposed in the literature for computing D-optimal designs [see e.g. Fedorov (1972), Wynn (1972), Silvey (1980), Pázman (1986) or the recent papers of Harman and Pronzato (2007), Mandal and Torsney (2006)].

In the present paper we concentrate on a class of multiplicative algorithms for computing D-optimal designs, which is indexed by real parameters, say β_r. For two special choices of the indexes β_r the algorithms are related to two algorithms proposed by Titterington (1976, 1978) and Silvey et al. (1978). In Section 2 we prove that the sequence of determinants of the corresponding designs is nondecreasing, and as a consequence the sequence of calculated designs converges to the D-optimal design. The monotonicity result uses particular thresholds in each step of the iterations, and some bounds for these thresholds are derived in Section 3. Some results of numerical comparisons are given in Section 4. In Section 5, some applications of multiplicative algorithms to the construction of optimal designs with respect to other optimality criteria are considered.

2. A class of multiplicative algorithms for calculating *D*-optimal designs

Let $w_1^{(0)}, \ldots, w_n^{(0)}$ denote a set of given initial weights for the design points $u_1, \ldots, u_n$, which defines the initial design $\xi^{(0)}$. For a design ξ with weights w_i at the points u_i, we define

$$d(u_i, \xi) = \frac{\partial}{\partial w_i} \log\det M(\xi) = u_i^{\mathrm{T}} M^{-1}(\xi) u_i \quad (i = 1, \ldots, n).$$

We study the class of multiplicative algorithms for calculating D-optimal designs which is defined recursively by updating the weights in each step as follows:

$$w_i^{(r+1)} = w_i^{(r)} \frac{d(u_i, \xi^{(r)}) - \beta_r}{m - \beta_r}, \quad \beta_r \in \mathbb{R}. \tag{2}$$

Here $r = 0, 1, \ldots$ is the iteration number and β_r $(r = 0, 1, \ldots)$ are real-valued parameters that can vary in each iteration. For an initial design $\xi^{(0)}$, the iterations (2) produce a sequence of designs $\{\xi^{(r)}\}_{r=0,1,\ldots}$.

The choices of constants $\beta_r = 0$ and $\beta_r = 1$ in algorithm (2) give two procedures considered by Titterington (1976, 1978), who discussed a special case of model (1), where the first component of the vector u in the regression model (1) is constant, i.e. $u = (1, v^{\mathrm{T}})^{\mathrm{T}}$, $v \in \mathbb{R}^{m-1}$. This author established that in the case $\beta_r = 0$ algorithm (2) yields a sequence of determinants $\{\det M(\xi^{(r)})\}$, which is nondecreasing and converges to the D-optimal design. A more detailed proof of the same result is given by Pázman (1986). Based on numerical experiments Silvey et al. (1978) and Titterington (1976, 1978) noticed that for $\beta_r = 1$ the algorithm converges substantially faster than for $\beta_r = 0$ and conjectured that the sequence of determinants for algorithm (2) with $\beta_r = 1$ is also nondecreasing. There is a vast literature where the rates of convergence of algorithm (2) with $\beta_r \in \{0, 1\}$ and related algorithms are numerically studied and the monotonicity conjecture for the case of $\beta_r = 1$ is numerically verified, see e.g. Pázman (1986), Pukelsheim and Torsney (1991), Torsney and Mandal (2001), Pronzato (2003), Harman and Pronzato (2007), Torsney (2007) and Pronzato et al. (2000), p.155.

Theorem 1 does not give a proof of the monotonicity conjecture in the case $\beta_r = 1$, but establishes the monotonicity of the determinants $\{\det M(\xi^{(r)})\}$ for certain positive values of β_r. For a precise formulation of the statement, we define for the design $\xi^{(r)}$ in the rth step the quantity

$$\beta^{(r)} = \min_{u_i \in \mathcal{U}} d(u_i, \xi^{(r)}),$$

which turns out to be essential in the following discussion. Note that we must always have $\beta_r \leq \beta^{(r)}$ as otherwise at least one weight in the updated design $\xi^{(r+1)}$ becomes negative.

Theorem 1. *Let $\{\xi^{(r)}\}_{r=0,1,\ldots}$ be a sequence of designs obtained by the recurrent formula (2), where in the rth step the parameter β_r is chosen as any number such that*

$$-\infty < \beta_r \leq \beta^{(r)}/2, \tag{3}$$

then the sequence of determinants $\{\det M(\xi^{(r)})\}_{r=0,1,\ldots}$ is nondecreasing and the sequence of designs $\{\xi^{(r)}\}$ converges to the D-optimal design ξ^*.

The proof (generalizing the proof of Pázman (1986) valid for $\beta_r = 0$) is based on several auxiliary lemmas.

Lemma 1. *Let $\zeta_1, \ldots, \zeta_m$ be i.i.d. random variables with values in $\mathcal{U} = \{u_1, \ldots, u_n\}$ and let α be some positive constant. Let also $g : \mathcal{U}^m \to [0, \infty)$ denote a function such that $E(g|\zeta_i) \geq 2\alpha$ a.s. (almost surely), where $g = g(\zeta_1, \ldots, \zeta_m)$. Then*

$$E(g)[E(g) - \alpha]^m \leq E\left[g \prod_{i=1}^{m} (E(g|\zeta_i) - \alpha)\right]$$

and the equality is strict unless $E(g|\zeta_1) = \cdots = E(g|\zeta_m) = E(g)$ a.s.

Lemma 2. *Let $h : \mathcal{U}^m \to \mathbb{R}$ be the function defined by*

$$h(z_1, \ldots, z_m) = \frac{1}{m!} \det{}^2 F(z_1, \ldots, z_m), \tag{4}$$

where $z_i \in \mathcal{U}$ $(i = 1, \ldots, m)$ and ith column of the matrix $F(z_1, \ldots, z_m)$ equals z_i. Then, for a design with weights $w_1, \ldots, w_n$ at the points $u_1, \ldots, u_n$, the determinant of the information matrix $M(\xi) = \sum_{i=1}^n w_i u_i u_i^{\mathrm{T}}$ can be represented as

$$\det M(\xi) = E\, h(\zeta_1, \ldots, \zeta_m) = \sum_{i_1=1}^{n} \ldots \sum_{i_m=1}^{n} w_{i_1} \cdots w_{i_m} h(u_{i_1}, \ldots, u_{i_m}), \tag{5}$$

where $\zeta_1, \ldots, \zeta_m$ are i.i.d. random variables with distribution ξ.

Lemma 3. *For any $\beta \in \mathbb{R}$ $(\beta \neq m)$, any $u \in \mathcal{U}$ and any j $(1 \leq j \leq m)$ we have*

$$\frac{d(u, \xi) - \beta}{m - \beta} = \frac{E\, h(\zeta_j^u) - \alpha}{E\, h(\zeta) - \alpha}$$

where $\zeta = (\zeta_1, \ldots, \zeta_m)$, $\zeta_j^u = (\zeta_1, \ldots, \zeta_{j-1}, u, \zeta_{j+1}, \ldots, \zeta_m)$, $\alpha = \beta E\, h(\zeta)/m$, $h(\cdot)$ is defined in (4) and $\zeta_1, \ldots, \zeta_m$ are (as in Lemmas 1 and 2) i.i.d. random variables with distribution ξ.

Proof of Lemma 1. From the inequality $t - 1 \geq \log(t)$ we obtain

$$\prod_{i=1}^{m} \frac{E(g|\zeta_i) - \alpha}{E(g) - \alpha} - 1 \geq \sum_{i=1}^{m} [\log(E(g|\zeta_i) - \alpha) - \log(E(g) - \alpha)] \text{ a. s.}$$

Multiplying both sides by g and taking the expectation we obtain

$$\begin{aligned} E\left\{ g \prod_{i=1}^{m} \frac{E(g|\zeta_i) - \alpha}{E(g) - \alpha} \right\} - E(g) &\geq \sum_{i=1}^{m} [E(g \log(E(g|\zeta_i) - \alpha)) - E(g) \log(E(g) - \alpha)] \\ &= \sum_{i=1}^{m} [E[E(g|\zeta_i) \log(E(g|\zeta_i) - \alpha)] - E(g) \log(E(g) - \alpha)] \geq 0. \end{aligned}$$

The last inequality follows from Jensen's inequality $E\phi(t) \geq \phi(E(t))$ with the strictly convex function $\phi(t) = t \log(t - \alpha)$, $t \in [2\alpha, \infty)$. ■

Proof of Lemma 2. This is a consequence of the Binet–Cauchy formula [see Gantmacher (1959), page 9]. ■

Proof of Lemma 3. By the definition, for any j $(1 \leq j \leq m)$ and $u = u_i \in \mathcal{U}$ we have

$$d(u_i, \xi) = \frac{\partial \log \det M(\xi)}{\partial w_i} = \frac{1}{\det M(\xi)} \frac{\partial \det M(\xi)}{\partial w_i} = \frac{m\, E\, h(\zeta_j^{u_i})}{E\, h(\zeta)},$$

where we have used (5). Therefore

$$\frac{d(u, \xi)}{m} = \frac{E\, h(\zeta_j^u)}{E\, h(\zeta)},$$

which yields

$$\frac{d(u, \xi) - \beta}{m - \beta} = \frac{d(u, \xi)/m - \beta/m}{1 - \beta/m} = \frac{E\, h(\zeta_j^u) - \alpha}{E\, h(\zeta) - \alpha}$$

where $\alpha = \beta E\, h(\zeta)/m$. ■

Proof of Theorem 1. Let $\zeta_1, \ldots, \zeta_m$ be i.i.d. random variables with distribution $\xi^{(r)}$, $h = h(\zeta_1, \ldots, \zeta_m)$, where the function h is defined in (4), and $\alpha = \beta_r E\, h/m$. Note that $E\, h(\zeta_j^u) = E\,(h(\zeta)|\zeta_j = u)$ and that (3) implies $E(h|\zeta_i) \geq 2\alpha$ a.s. The last inequality allows us to apply Lemma 1 with $g = h$. Lemmas 1 and 3 yield

$$\det M(\xi^{(r+1)}) = E\left\{ h \prod_{i=1}^{m} \frac{E(h|\zeta_i) - \alpha}{E\, h - \alpha} \right\} \geq E\, h = \det M(\xi^{(r)}). \tag{6}$$

Note that to obtain the left equality in (6) we have used formulae (5) and (2). The inequality (6) becomes equality if and only if $\xi^{(r+1)} = \xi^{(r)}$ a.s.

This proves the monotonicity result. To prove convergence of the sequence $\{\xi^{(r)}\}$ to ξ^* we use standard arguments, see e.g. Pázman (1986, p. 142).

Since the sequence $\{\xi^{(r)}\}$ is bounded, there is at least one limiting point of $\{\xi^{(r)}\}$. Moreover, in view of the monotonicity of the sequence $\{\det M(\xi^{(r)})\}$ all limiting points of $\{\xi^{(r)}\}$ have the same determinant of the information matrix which is equal to $\lim_r \det M(\xi^{(r)})$. Let $\{\xi^{(r_j)}\}$ be a subsequence of the sequence $\{\xi^{(r)}\}$ converging to some design $\bar{\xi}$. Let $\{\xi^{(r_{j_s}-1)}\}$ also be a subsequence of $\{\xi^{(r_j-1)}\}$ converging to a design $\tilde{\xi}$. Denote by T the mapping defined by $\xi^{(r)} = T\xi^{(r-1)}$. From (2) we obtain that $T\tilde{\xi} = \bar{\xi}$. Using the facts that $\det M(\bar{\xi}) = \det M(\tilde{\xi})$ and that inequality (6) is equality if and only if $\xi^{(r+1)} = \xi^{(r)}$ a.s., we obtain $\bar{\xi} = T\tilde{\xi} = \tilde{\xi}$. The equivalence theorem (Kiefer and Wolfowitz, 1960) now implies that $\bar{\xi}$ is the D-optimal design as it is a unique fixed point of mapping T. ■

Remark 1 (*Sharpness of the Main Result*). Theorem 1 is sharp in the following sense. If we consider algorithm (2) with $\beta_r = \gamma\beta^{(r)}$ and fixed γ, then $\gamma^* = \frac{1}{2}$ is the largest possible value of γ such that algorithm (2) yields a monotonic sequence of determinants $\{\det M(\xi^{(r)})\}$ for any regression model, any $\mathcal{U}$ and any initial design (note that for some models and some initial designs the value of γ can be increased, see below).

Indeed, consider the regression model $y = \theta^{\mathrm{T}}u + \varepsilon$, where $\mathcal{U}$={u_0, u_1} = {$(1, 0)^{\mathrm{T}}, (1, 1)^{\mathrm{T}}$}, and let the weights of a two-point design $\xi^{(0)}$ be $w_0 = \frac{1}{2} + \varepsilon$ and $w_1 = \frac{1}{2} - \varepsilon$ for some small $\varepsilon > 0$. Assume that $\beta_r = (\frac{1}{2} + \delta)\beta^{(r)}$ for some small $\delta > 0$. It is easy to compute

$$\det M(\xi^{(1)}) - \det M(\xi^{(0)}) = \frac{8\varepsilon^2(1 + 2\varepsilon)}{(1 + 4\varepsilon - 2\delta)^2}\,(\varepsilon - \delta)$$

which is negative if $\delta > \varepsilon$.

Remark 2 (*Regression Model with Intercept*). Consider the regression model with intercept, where

$$u = (1, v^{\mathrm{T}})^{\mathrm{T}}, \quad v \in \mathbb{R}^{m-1}.$$

It is easy to show [see Titterington (1978)] that in this case the function $d(u_i, \xi)$ can be written as $d(u_i, \xi) = d_c(v_i, \xi) + 1$, where

$$d_c(v_i, \xi) = (v_i - \bar{v})^{\mathrm{T}} M_c^{-1}(v_i - \bar{v}), \tag{7}$$

$$M_c = \sum_{i=1}^{n} w_i(v_i - \bar{v})(v_i - \bar{v})^{\mathrm{T}}, \quad \bar{v} = \sum_{i=1}^{n} w_i v_i.$$

Therefore, $d(u_i, \xi) \geq 1$ for any $u_i \in \mathcal{U}$. As a consequence of Theorem 1, the algorithm in the model with intercept is monotonic for any $\beta_r \leq \frac{1}{2}$. Note that for some models and designs $\xi^{(r)}$, the values of $\beta_r = \beta^{(r)}/2$ may be much larger than 1.

Remark 3 (*Minimal Volume Ellipsoids*). As demonstrated by many authors [see e.g. Titterington (1975, 1978) and Pronzato (2003) the multiplicative algorithms for finding D-optimal designs play an important role in the construction of a minimal volume ellipsoid containing a set of points

$$\mathcal{V} = \{v_1, \ldots, v_n\} \in \mathbb{R}^{m-1}. \tag{8}$$

Indeed, let the data set consists of the points (8). Then the minimal covering ellipsoid for the set $\mathcal{V}$ is of the form

$$\mathcal{E}(\mathcal{V}) = \{v : d_c(v, \xi) \leq m - 1\}$$

where $d_c(v, \xi)$ is defined in (7). In other terms this ellipsoid has the form

$$\{u : d(u, \xi) \leq m\}.$$

By the equivalence theory of Kiefer and Wolfowitz (1960) we have that the point v_i has a positive weight w_i only if it lies on the surface of the ellipsoid $\mathcal{E}(\mathcal{V})$. Thus, the points that lie on the surface of the minimal volume ellipsoid can be found from the solution of the D-optimal design problem on a finite set. Therefore, the algorithms whose convergence have been established in Theorem 1 can be considered as algorithms for the construction of the minimal volume ellipsoids containing given sets of points. As illustrated below on some numerical examples, these algorithms may have much faster convergence than the algorithms considered in Titterington (1976, 1978).

Remark 4 (*Arbitrary Design Space*). In the case of an arbitrary (i.e. not necessarily discrete) design space $\mathcal{U}$, algorithm (2) becomes

$$\mathrm{d}\xi^{(r+1)}(u) = \frac{d(u, \xi^{(r)}) - \beta_r}{m - \beta_r}\,\mathrm{d}\xi^{(r)}(u), \quad \beta_r \in \mathbb{R}, u \in \mathcal{U}. \tag{9}$$

In particular, if the initial design $\xi^{(0)}$ has a density, then all designs $\xi^{(r)}$ have densities $p_r(u)$, that is $d\xi^{(r)}(u) = p_r(u)\mathrm{d}u$, and the updating formula (9) has the form

$$p_{r+1}(u) = \frac{d(u, \xi^{(r)}) - \beta_r}{m - \beta_r}\, p_r(u), \quad \beta_r \in \mathbb{R}, u \in \mathcal{U}.$$

Theorem 1 can be generalized to establish monotonicity of determinants for iterations (9).

3. A lower bound for the function *d*

In this section we derive a lower bound for the function d. Recall the definition of the Elfving set corresponding to the linear regression model (1), i.e.

$$\mathcal{G} = \text{conv}\{u_1, \ldots, u_n, -u_1, \ldots, -u_n\}$$

[see Elfving (1952) or Pukelsheim (1993)], where conv(A) denotes the convex hull of a set A. Note that $\mathcal{G}$ is a nonempty convex set containing the origin if there are m linearly independent vectors among $u_1, \ldots, u_n$. For a vector $u \in \mathbb{R}^m \setminus \{0\}$, define γ_u as the positive number such that the scaled vector $\gamma_u u$ is a boundary point of $\mathcal{G}$. Note that γ_u is uniquely defined and $\gamma_u \geq 1$ for any $u \in \mathcal{G}$. Define also

$$\beta^* = \min_{u \in \mathcal{U}} \frac{1}{\gamma_u^2}.$$

Theorem 2. *For any model (1), we have*

(i) $\beta^* \leq d(u, \xi)$ *for any* $u \in \mathcal{U}$ *and any* ξ*;*
(ii) $0 \leq \beta^* \leq 1$*;*
(iii) $\beta^* = 1$ *if and only if all* $u_i \in \mathcal{U}$ *belong to the boundary of the Elfving set for* $\mathcal{U}$*.*

Proof of Theorem 2. For any $u \in \mathbb{R}^m \setminus \{0\}$, let ξ_u^* be a c-optimal design (for the vector $c = u$). Then it follows from the optimality of the design ξ_u^* that for any $u \in \mathcal{U}$ and any design ξ such that $c^{\mathrm{T}}\theta$ is estimable

$$d(u, \xi) = u^{\mathrm{T}} M^{-}(\xi) u \geq u^{\mathrm{T}} M^{-}(\xi_u^*) u,$$

where A^- denotes a generalized inverse of the matrix A. Moreover, the right-hand side of this inequality is known (Pukelsheim, 1993) to be

$$u^{\mathrm{T}} M^{-}(\xi_u^*) u = \frac{1}{\gamma_u^2}.$$

This proves assertion (i). Assertion (ii) follows from the definition of β^* and the fact that $\gamma_u \geq 1$ for any $u \in \mathcal{U}$. For a given $u \in \mathcal{U}$, $\gamma_u = 1$ if u is a boundary point of the Elfving set $\mathcal{G}$. Therefore, $\beta^* = 1$ if and only if all $u_i \in \mathcal{U}$ are boundary points of $\mathcal{G}$, which proves (iii). ■

Remark 5 (*Model with Intercept*). For models with a constant term we always have $\beta^* = 1$ as all points $u_i = (1, v_i^{\mathrm{T}})^{\mathrm{T}}$ are located at the boundary of the corresponding Elfving set.

4. Numerical comparisons

In this section, we present a few numerical comparisons of the algorithms. We discuss the performance of the algorithms for several polynomial, exponential and rational regression models. To be precise, we consider the regression model

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

where $f(x) = (f_0(x), \ldots f_{m-1}(x))^{\mathrm{T}}$ is the vector of regression functions and the explanatory variable x varies in a finite set, say $\mathcal{X} = \{x_1, \ldots, x_n\}$. In this case, we have $\mathcal{U} = \{f(x_1), \ldots, f(x_n)\}$ and the regression model can be written in the form (1) with $u = f(x)$. For the polynomial regression model we choose

$$f(x) = (1, x, \ldots, x^{m-1})^{\mathrm{T}}, \quad m = 3, 4, 5, 6,$$

while the exponential and rational models are given by

$$f(x) = (1, \mathrm{e}^{-x}, x\mathrm{e}^{-x})^{\mathrm{T}},$$
$$f(x) = (\mathrm{e}^{-x}, x\mathrm{e}^{-x}, \mathrm{e}^{-2x}, x\mathrm{e}^{-2x})^{\mathrm{T}},$$
$$f(x) = (1, \mathrm{e}^{-x}, x\mathrm{e}^{-x}, \mathrm{e}^{-2x}, x\mathrm{e}^{-2x})^{\mathrm{T}},$$
$$f(x) = (1, 1/(1+x), 1/(1+x)^2)^{\mathrm{T}}.$$

We consider two design spaces that correspond to Tables 1 and 2, respectively. Specifically, we consider

$$\mathcal{X}_1 = \{4i/19 \mid i = 0, \ldots, 19\} \quad \text{and} \quad \mathcal{X}_2 = \{4i/39 \mid i = 0, \ldots, 39\}.$$

Table 1
The number of iterations required to achieve precision (10) by algorithm (2) with parameters $\beta_r = \gamma\beta^{(r)}$ and by Titterington's algorithm with $\beta_r = 1$. The initial design $\xi^{(0)}$ is a uniform distribution on the set $\mathcal{X}_1 = \{4i/19 \mid i = 0, \ldots, 19\}$.

	Algorithm (2) with $\beta_r = \gamma\beta^{(r)}$									$\beta_r = 1$
γ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
$(1, x, x^2)$	104	97	91	84	78	71	65	58		69
$(1, x, x^2, x^3)$	130	121	113	104	96	88	79	71		98
$(1, x, \ldots, x^4)$	82	77	72	67	61	56	51	45		66
$(1, x, \ldots, x^5)$	96	89	82	75	68	61	53			80
$(1, e^{-x}, xe^{-x})$	131	123	115	108	100	92	84	76		90
$(1, 1/(1+x), 1/(1+x)^2)$	105	98	92	85	79	73	66	60		71
$(e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	221	208	196	183	170	158	145	133		167
$(1, e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	136	127	118	109	100	91	83	74		109

Table 2
The number of iterations required to achieve precision (10) by algorithm (2) with parameters $\beta_r = \gamma\beta^{(r)}$ and by Titterington's algorithm with $\beta_r = 1$. The initial design $\xi^{(0)}$ is a uniform distribution on the set $\mathcal{X}_2 = \{4i/39 \mid i = 0, \ldots, 39\}$.

	Algorithm (2) with $\beta_r = \gamma\beta^{(r)}$									$\beta_r = 1$
γ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	
$(1, x, x^2)$	250	235	219	204	188	172	157	141		167
$(1, x, x^2, x^3)$	329	308	287	266	244	223	202	181		247
$(1, x, \ldots, x^4)$	235	219	204	188	173	157	142	127		188
$(1, x, \ldots, x^5)$	281	262	244	226	207	189	170	152		234
$(1, e^{-x}, xe^{-x})$	294	276	258	239	221	202	184	166		197
$(1, 1/(1+x), 1/(1+x)^2)$	136	128	120	111	103	94	86	77		91
$(e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	404	382	359	337	314	291	269	246	224	304
$(1, e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	213	199	185	171	157	143	130	116		171

In all cases a uniform distribution on the design space $\mathcal{X}$ was used as starting design $\xi^{(0)}$, and the iterations of algorithm (2) were performed until the precision

$$\max_{i=1,\ldots,n} d(u_i, \xi^{(r)}) \leq 1.001m \tag{10}$$

was reached. Note that the expression $m/\max_{i=1}^{n} d(u_i, \xi^{(r)})$ provides a lower point for the D-efficiency of the design $\xi^{(r)}$ [see Dette (1996)]. Thus with this stopping rule the D-efficiency of the design $\xi^{(r)}$ is at least $(1.001)^{-1}$ and other constants would give different lower bounds for the D-efficiency. In Tables 1 and 2 we present the number of iterations required to reach this precision by algorithm (2) with $\beta_r = \gamma\beta^{(r)}$, for various values of γ. The tables also contain the results for Titterington's algorithm, which uses constant parameter $\beta_r = 1$ (see the right columns in the tables). The empty space in the table indicates that the corresponding algorithm did not converge to the optimal design (note that the convergence is only proved for $\gamma \leq 0.5$).

Both tables show very similar results. The performance of algorithm (2) is improved if larger values are used for the parameter γ. The worst case corresponds to the choice $\beta_r = 0$, and this is improved by all other methods. If γ is small, the algorithm with $\beta_r = 1$ is still better than the method proposed in this paper. However this picture is changing if values $\gamma \geq 0.5$ are used in the procedure. Note that it follows from the proof of Theorem 1 that the sequence of determinants generated by algorithm (2) may still be nondecreasing for values of β_r, which are slightly larger than $\beta^{(r)}/2$ and for this reason our numerical comparison also includes the cases $\gamma = 0.6$, 0.7 and 0.8 (see Tables 1 and 2). If $\gamma \geq 0.5$, the algorithm (2) is at least comparable to the case $\beta_r = 1$, and in many cases it yields a substantially smaller number of iterations for achieving the desired precision. Since in practical computation we can not risk to use the algorithm with $\gamma > 0.5$, we recommend the choice $\gamma = 0.5$.

Note that although the algorithm with $\beta_r = \gamma\beta^{(r)}$ requires certain time for the calculation of the minimum $\beta^{(r)} = \min_i d(u_i, \xi^{(r)})$ (compared with the algorithms which use $\beta_r = 0$ and $\beta_r = 1$), this time is typically a very small part of the total time needed to perform an iteration. Note also that the convergence of the algorithm with $\beta_r = 1$ is still not proved. We finally mention that a further improvement of the rate of convergence of the algorithm can be obtained if nonoptimal points are removed at each iteration using the rule developed in Harman and Pronzato (2007).

5. Other optimality criteria

5.1. Bayesian D-optimal designs

In this section we briefly discuss an extension of the multiplicative algorithm to the Bayesian D-optimal design problem. To be precise, consider the nonlinear regression model

$$Y = \eta(x, \theta) + \varepsilon,$$

Table 3
The number of iterations required to achieve precision (12) by algorithm (11) with parameters $\beta_r = \gamma\beta^{(r)}$ and distribution $\{0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3; 1/7, \ldots, 1/7\}$ with respect to parameter θ_2. The initial design $\xi^{(0)}$ is a uniform distribution on the set $\mathcal{X}_3 = \{3i/19 \mid i = 0, \ldots, 19\}$.

	Algorithm (11) with $\beta_r = \gamma\beta^{(r)}$								$\beta_r = 1$
γ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
$\theta_1 + \theta_3 e^{-\theta_2 x}$	178	167	156	145	133	122	111	100	120
$\theta_1 + \theta_3/(\theta_2 + x)$	147	138	129	120	110	101	92	83	98
$\theta_1 e^{-\theta_2 x} + \theta_3 e^{-2x} + \theta_4 e^{-2x}$	322	296	270	244	218	192	165		242
$\theta_1 + \theta_3 e^{-\theta_2 x} + \theta_4 e^{-2x} + \theta_5 e^{-2x}$	101	95	88	81	75	68	61	55	81

where η is a known function, $\theta \in \Theta \subset \mathbb{R}^m$ denotes the unknown parameter and the explanatory variable x varies in the finite design space $\mathcal{X} = \{x_1, \ldots, x_n\}$. Under the assumption of a normally distributed homoscedastic error it was shown by Jennrich (1969) that asymptotically the covariance matrix of the least squares estimate for the parameter θ is proportional to the matrix $M^{-1}(\xi, \theta)$, where ξ is the given design,

$$M(\xi, \theta) = \int_{\mathcal{X}} f(x, \theta) f^{\mathrm{T}}(x, \theta) \mathrm{d}\xi(x),$$

and $f(x, \theta) = \frac{\partial}{\partial\theta}\eta(x, \theta)$ is the gradient of η with respect to θ. A Bayesian D-optimal design maximizes

$$\Phi_\pi(\xi) = \int_{\Theta} \log\det M(\xi, \theta)\pi(\mathrm{d}\theta)$$

where π denotes a given prior distribution on the parameter space Θ [see e.g. Chaloner and Larntz (1989) or Chaloner and Verdinelli (1995)]. Define

$$d_\pi(x, \xi) = \int_{\Theta} f^{\mathrm{T}}(x, \theta) M^{-1}(\xi, \theta) f(x, \theta)\pi(\mathrm{d}\theta)$$

and note that a design ξ^* is Bayesian D-optimal if and only if the inequality

$$d_\pi(x, \xi^*) \leq m$$

holds for all $x \in \mathcal{X}$. We consider the multiplicative algorithm

$$w_i^{(r+1)} = w_i^{(r)} \frac{d_\pi(x_i, \xi^{(r)}) - \beta_r}{m - \beta_r}, \quad i = 1, \ldots, n, \tag{11}$$

where the procedure is terminated if

$$\max_{i=1,\ldots,n} d_\pi(x_i, \xi^{(r)}) \leq 1.001m. \tag{12}$$

Some numerical results for algorithm (11) are given in Table 3. We can see that the performance of algorithm (11) is similar to the performance of algorithm (2).

Based on extensive numerical calculations we conjecture that Theorem 1 can be extended from the D-optimality criterion to the Bayesian D-optimality criterion; that is, the sequence of designs defined by (11) with $\beta_r \leq \frac{1}{2}\min_{i=1,\ldots,n} d_\pi(x_i, \xi^{(r)})$ yields a nondecreasing sequence $\{\Phi_\pi(\xi^{(r)})\}$.

5.2. A-, E- and c-optimal designs

Let us finally discuss an extension of the multiplicative algorithm to the A-, E- and c-optimal design problems. Consider a general (differentiable) optimality criterion Φ such that Φ-optimal design either maximizes $\Phi(M(\xi))$. It is often easier to consider Ψ-optimal design which minimizes $\Psi(M^{-1}(\xi))$ with respect to ξ. The nonnegative function $d(u_i, \xi)$ is then generalized either to

$$\phi(u_i, \xi) = \frac{\partial\Phi(M(\xi))}{\partial w_i} = u_i^{\mathrm{T}} \overset{\circ}{\Phi}(\xi) u_i$$

or

$$\phi(u_i, \xi) = -\frac{\partial\Psi(M^{-1}(\xi))}{\partial w_i} = u_i^{\mathrm{T}} M^{-1}(\xi) \overset{\circ}{\Psi}(\xi) M^{-1}(\xi) u_i$$

where

$$\overset{\circ}{\Phi}(\xi) = \left.\frac{\partial\Phi(M)}{\partial M}\right|_{M=M(\xi)}, \qquad \overset{\circ}{\Psi}(\xi) = \left.\frac{\partial\Psi(M^{-1})}{\partial M^{-1}}\right|_{M=M(\xi)}.$$

Table 4
The number of iterations required to achieve precision (14) by algorithm (13) for E-optimality with parameters $\beta_r = (1-\gamma)b(\xi^{(r)})$. The initial design $\xi^{(0)}$ is a uniform distribution on the set $\mathcal{X}_3 = \{3i/19 \mid i = 0, \ldots, 19\}$.

	Algorithm (13) with $\beta_r = (1-\gamma)b(\xi^{(r)})$									
γ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$(1, x, x^2)$	100	95	90	85	80	75	70	65		
$(1, x, x^2, x^3)$	129	122	116	110	103	97	90	84	78	
$(1, x, \ldots, x^4)$	51	48	46	43	41	38	35	33	31	
$(1, x, \ldots, x^5)$	215	204	194	183	172	162	151	141	130	120
$(1, e^{-x}, xe^{-x})$	265	252	239	226	213	200	187	174		
$(1, 1/(1+x), 1/(1+x)^2)$	115	109	103	98	92	86	80	75	69	62
$(e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	493	469	444	419	395	370	346	321	297	272
$(1, e^{-x}, xe^{-x}, e^{-2x}, xe^{-2x})$	90	86	81	77	72	68	63	59	54	50

For the A-optimality criterion $\Psi(M^{-1}(\xi)) = \mathrm{tr}(M^{-1}(\xi))$, it follows $\overset{\circ}{\Psi}(\xi) = I_m$ and the function ϕ has the form

$$\phi_A(u_i, \xi) = u_i^{\mathrm{T}} M^{-2}(\xi) u_i.$$

If the multiplicity of the minimal eigenvalue of the matrix $M(\xi)$ equals 1, then for the E-optimality criterion $\Phi(M(\xi)) = \lambda_{\min}(M(\xi))$ (which is in general not differentiable) the function ϕ has the form

$$\phi_E(u_i, \xi) = (p^{\mathrm{T}} u_i)^2,$$

where p is a normalized eigenvector corresponding to the minimal eigenvalue of $M(\xi)$. If the matrix $M(\xi)$ has rank m, the function ϕ for the c-optimality criterion $\Psi(M^{-1}(\xi)) = c^{\mathrm{T}} M^{-1}(\xi) c$ is given by

$$\phi_c(u_i, \xi) = (c^{\mathrm{T}} M^{-1}(\xi) u_i)^2.$$

For D-, A-, E- and c-optimality we consider an algorithm in the following form

$$w_i^{(r+1)} = w_i^{(r)} \frac{\phi(u_i, \xi^{(r)}) + \beta_r}{b(\xi^{(r)}) + \beta_r}, \quad \beta_r \in \mathbb{R} \tag{13}$$

where $b(\xi) = \mathrm{tr}\, M(\xi)\, \overset{\circ}{\Phi}(\xi)$ or $b(\xi) = \mathrm{tr}\, M^{-1}(\xi)\, \overset{\circ}{\Psi}(\xi)$. Note that sum of weights at the next iteration equals 1 as

$$b(\xi^{(r)}) = \sum_j w_j^{(r)} \phi(u_j, \xi^{(r)})$$

and that for the D-criterion (13) reduces to the recursive relation defined in (2).

Numerical calculations show that the algorithm (13) with $\beta_r = 0$ is generally not monotonic (that is, the sequence $\{\Phi(M(\xi^{(r)}))\}$ is not a monotone sequence) for the A-, E- and c-criteria, in contrast to the case of D-optimality. We therefore need to use positive values of β_r in algorithm (13). We conjecture that for A-, E- and c-optimality, the sequence of designs $\{\xi^{(r)}\}$ obtained by the recurrent formula (13) with $\beta_r \geq \frac{1}{2} b(\xi^{(r)})$ yields a monotonic sequence $\{\Phi(M(\xi^{(r)}))\}$.

In numerical studies, we use $\beta_r = (1-\gamma)b(\xi^{(r)})$ for different values of γ, $0 \leq \gamma < 1$. We run the iterations of the algorithm (13) until the precision

$$\max_{i=1,\ldots,n} \phi(u_i, \xi^{(r)}) \leq 1.001\, b(\xi^{(r)}) \tag{14}$$

is achieved. Some results are given in Table 4 and demonstrate that multiplicative algorithms can also be applied to other optimality criteria.

Acknowledgements

The work was partly supported by the Deutsche Forschungsgemeinschaft (SFB 475: Komplexitätsreduktion in multivariaten Datenstrukturen, Teilprojekt A2). The work of H. Dette was also supported in part by a NIH grant award IR01GM072876:01A1. The authors are very grateful to Radoslav Harman and to two unknown referees who carefully checked the proofs and made a few very useful comments on an earlier version of this paper. The authors are also grateful to M. Stein who typed parts of this paper with considerable technical expertise.

References

Chaloner, K., Larntz, K., 1989. Optimal Bayesian design applied to logistic regression experiments. J. Statist. Plann. Inference 21 (2), 191–208.
Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statist. Sci. 10 (3), 273–304.

Dette, H., 1996. Lower bounds for efficiencies with applications. In: Brunner, E., Denker, M. (Eds.), Research Developments in Probability and Statistics: Festschrift in Honor of Madan L. Puri on the Occasion of his 65th Birthday. VSP Utrecht, The Netherlands, pp. 111–124.
Elfving, G., 1952. Optimum allocation in linear regression theory. Ann. Math. Statist. 23, 255–262.
Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press, New York.
Gantmacher, F.R., 1959. The Theory of Matrices, volume 1. Chelsea, New York.
Harman, R., Pronzato, L., 2007. Improvements on removing nonoptimal support points in *D*-optimum design algorithms. Statist. Probab. Lett. 77 (1), 90–94.
Jennrich, R.I., 1969. Asymptotic properties of non-linear least squares estimators. Ann. Math. Statist. 40, 633–643.
Kiefer, J., 1974. General equivalence theory for optimum designs (approximate theory). Ann. Statist. 2 (5), 849–879.
Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canad. J. Math. 12, 363–366.
Mandal, S., Torsney, B., 2006. Construction of optimal designs using a clustering approach. J. Statist. Plann. Inference 136, 1120–1134.
Pázman, A., 1986. Foundations of Optimum Experimental Design. Reidel, Dordrecht.
Pronzato, L., 2003. Removing non-optimal support points in *D*-optimum design algorithms. Statist. Probab. Lett. 63 (3), 223–228.
Pronzato, L., Wynn, H.P., Zhigljavsky, A.A., 2000. Dynamical Search. Chapman & Hall/CRC, Boca Raton, FL.
Pukelsheim, F., 1993. Optimal Design of Experiments. John Wiley & Sons Inc, New York.
Pukelsheim, F., Rieder, S., 1992. Efficient rounding of approximate designs. Biometrika 79 (4), 763–770.
Pukelsheim, F., Torsney, B., 1991. Optimal weights for experimental designs on linearly independent support points. Ann. Statist. 19 (3), 1614–1625.
Silvey, S.D., 1980. Optimal Design. Chapman & Hall, London.
Silvey, S.D., Titterington, D.M., Torsney, B., 1978. An algorithm for optimal designs on a finite design space. Commun. Stat. Theory Methods 14, 1379–1389.
Titterington, D.M., 1975. Optimal design: Some geometrical aspects of *D*-optimality. Biometrika 62 (2), 313–320.
Titterington, D.M., 1976. Algorithms for computing D-optimal design on finite design spaces. In Proc. of the 1976 Conf. on Information Science and Systems, John Hopkins University, 3, pp. 213–216.
Titterington, D.M., 1978. Estimation of correlation coefficients by ellipsoidal trimming. Appl. Stat. 27 (3), 227–234.
Torsney, B., 2007. W-iterations and ripples therefrom. In: Pronzato, L., Zhigljavsky, A. (Eds.), Optimal Design and Related Areas in Optimization and Statistics. Springer-Verlag, N.Y, pp. 1–12.
Torsney, B., Mandal, S., 2001. Construction of constrained optimal designs. In: Atkinson, A.C., Bogacka, B., Zhigljavsky, A. (Eds.), Optimum Design 2000. Kluwer Acad. Publ, Dordrecht, pp. 141–152.
Wynn, H.P., 1972. Results in the theory and construction of *D*-optimum experimental designs. J. Roy. Statist. Soc. Ser. B 34, 133–147.