The Annals of Statistics
1984, Vol. 12, No. 1, 283–300

OPTIMAL BAYESIAN EXPERIMENTAL DESIGN FOR LINEAR MODELS[1]

BY KATHRYN CHALONER

University of Minnesota

Optimal Bayesian experimental designs for estimation and prediction in linear models are discussed. The designs are optimal for estimating a linear combination of the regression parameters $\mathbf{c}^T\theta$ or prediction at a point where the expected response is $\mathbf{c}^T\theta$ under squared error loss. A distribution on $\mathbf{c}$ is introduced to represent the interest in particular linear combinations of the parameters. In the usual notation for linear models minimizing the preposterior expected loss leads to minimizing the quantity $\text{tr}\psi(R + XX^T)^{-1}$. The matrix ψ is defined to be $E(\mathbf{c}\mathbf{c}^T)$ and the matrix R is the prior precision matrix of θ. A geometric interpretation of the optimal designs is given which leads to a parallel of Elfving's theorem for $\mathbf{c}$-optimality. A bound is given for the minimum number of points at which it is necessary to take observations. Some examples of optimal Bayesian designs are given and optimal designs for prediction in polynomial regression are derived. The optimality of rounding non-integer designs to integer designs is discussed.

1. Introduction. Optimal experimental designs for classical linear models have received and continue to receive considerable attention in the statistical literature. Much of the pioneering work in this area is due to Kiefer, for example in Kiefer (1959, 1961, 1974) and Kiefer and Wolfowitz (1959, 1960, 1965). Optimal experimental design is discussed at length in a book by Fedorov (1972) and more recently in Silvey (1980). Optimal designs have not been extensively studied however, in a Bayesian framework. An optimal Bayesian design depends not only on what functions of the parameters are to be estimated or at what values of the independent variables prediction is required, but also on the prior distribution of the regression parameters.

Optimal experimental designs are derived here for estimation and prediction in Bayesian linear models. The designs derived are optimal under expected squared error loss and the assumptions of normality, independence, and homoscedasticity usually made in linear models.

We will assume, as usual, that we can observe a vector $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ such that

$$\mathbf{y} = X^T\theta + \mathbf{e}$$

where $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ is the $k \times n$ design matrix and each $\mathbf{x}_i$ is a k-dimensional column vector, $\theta = (\theta_1, \theta_2, \cdots, \theta_k)^T$ is a vector of k unknown parameters and $\mathbf{e} \mid \tau \sim N(\mathbf{0}, \tau I)$ is the n-dimensional random vector of observa-

Received December 1981; revised August 1983.
[1] Research supported in part by NSF grant SES-7906386.
AMS 1980 *subject classifications.* Primary 62K05; secondary 62F15.
Key words and phrases. Optimal design, linear models, polynomial regression.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

tion errors having a normal distribution with mean vector $\mathbf{0} = (0, 0, \ldots, 0)^T$ and precision matrix τI. The parameter τ can be assumed to be either known or unknown and I is the $n \times n$ identity matrix. We assume that the prior distribution of θ and τ is such that the conditional distribution of θ given τ is $N(\theta_0, \tau R)$, where R is a specified positive definite $k \times k$ matrix. The distribution of τ is arbitrary, although it will be necesary to assume that $E(\tau^{-1})$ is finite in order to ensure that the risk associated with optimal designs is finite. In particular, τ may be known or may have a gamma distribution which is a conjugate prior distribution. In any case, the posterior conditional distribution of θ given τ, that is $p(\theta \mid \tau, \mathbf{y})$, is normal with mean $\theta_1 = (R + XX^T)^{-1}(X\mathbf{y} + R\theta_0)$ and precision matrix $\tau(R + XX^T)$. If we are interested in a particular combination $\mathbf{c}^T\theta$ of the θ_i's and squared error loss is appropriate, the best point estimate is the posterior mean $\mathbf{c}^T\theta_1$. The posterior risk is the expected variance of $\mathbf{c}^T\theta_1$, that is $\mathbf{c}^T(R + XX)^{-1}\mathbf{c}E_{\tau \mid y}(\tau^{-1})$, where $E_{\tau \mid y}(\tau^{-1})$ is the posterior mean of τ^{-1}. Thus the preposterior risk in estimating $\mathbf{c}^T\theta$ is just $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}E(\tau^{-1})$ where $E(\tau^{-1})$ is the prior mean of τ^{-1}. Hence, assuming that $E(\tau^{-1})$ is finite, the optimal design on n points, i.e., the optimal choice of the $k \times n$ matrix X, to estimate $\mathbf{c}^T\theta$ would be an X such that $XX^T = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$ minimizes $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$. We are assuming that the $\mathbf{x}_i$'s are not stochastic and can be chosen by the experimenter from some specified set $\mathscr{X}$.

A more typical situation would require estimation of $\mathbf{c}^T\theta$ for more than one vector $\mathbf{c}$. This situation could be represented by putting a probability measure μ on $\mathbf{c}$. For example, if the experimenter wishes to estimate each θ_i separately and each is equally important, μ might put weight k^{-1} in the direction of unit vectors along each coordinate axis. With a probability measure on $\mathbf{c}$ it would be appropriate to choose X to minimize the expected preposterior risk, where the expectation is taken with respect to μ. Thus our criterion for optimality is the minimization of $E_\mu[\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}]$, which can also be written as $\operatorname{tr} \psi(R + XX^T)^{-1}$, with $\psi = E_\mu(\mathbf{c}\,\mathbf{c}^T)$. The expected risk is then $\operatorname{tr} \psi(R + XX^T)^{-1}E(\tau^{-1})$. This criterion which we call ψ-optimality was proposed and discussed in Duncan and DeGroot (1976). Note that ψ-optimality could also be derived by assuming a quadratic loss of $(\theta - \hat{\theta})^T\psi(\theta - \hat{\theta})$.

It is also true that ψ-optimal designs are optimal designs for prediction at certain points. If the value of the dependent variable Y is to be predicted at a point where its expected value is $\mathbf{c}^T\theta$, then the best prediction under squared error loss is again the posterior mean of $\mathbf{c}^T\theta$, and the expected risk is $(1 + \mathbf{c}^T(R + XX^T)^{-1}\mathbf{c})E(\tau^{-1})$. If we were to put a probability measure μ over $\mathbf{c}$, then the expected risk is $(1 + E_\mu[\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}])E(\tau^{-1})$, which can be written as $(1 + \operatorname{tr} \psi(R + XX^T)^{-1})E(\tau^{-1})$, with $\psi = E_\mu\mathbf{c}\mathbf{c}^T$. Thus the optimal design is again the XX^T minimizing $\operatorname{tr} \psi(R + XX^T)^{-1}$.

ψ-optimality has also been referred to as Bayes A-optimality. The criterion complies with the view stated by Lindley (1968) that the design of an experiment should depend on the use that is to be made of the information obtained. If the purpose of the experiment is not estimation or prediction then ψ-optimality may not be appropriate.

A ψ-optimal design can also be interpreted as the solution to several non-

Bayesian problems. Suppose the experimenter is committed to a design matrix X_0 but can also choose n further observations. Also suppose that for some given vector $\mathbf{c}$, the experimenter wishes to minimize the variance of $\mathbf{c}^T\hat{\theta}$, the least squares estimate of $\mathbf{c}^T\theta$. The appropriate criterion is to minimize $\tau^{-1}\mathbf{c}^T(X_0X_0^T + XX^T)^{-1}\mathbf{c}$. If $X_0X_0^T$ is of full rank this is exactly the problem of minimizing $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$. This is discussed in Silvey (1969). In the context of designs for estimating response surfaces Kiefer (1973) suggested minimizing the criterion $\mathrm{tr}\psi(R + XX^T)^{-1}$ when bias is believed to be present. The matrices ψ and R however, are derived from different considerations than those presented here. Finally the criterion of ψ-optimality has also been shown to be appropriate in situations where certain kinds of prior knowledge are accounted for using a non-Bayesian approach. Näther and Pilz (1980) detail these situations. One example is when θ is constrained to lie in a region $(\theta - \theta_0)^T R(\theta - \theta_0) \leq \tau^{-1}$ and minimax linear estimation is used (see eg. Rao, 1976). Another example is the mixed estimation procedure of Theil (see eg. Theil, 1971, pages 346).

The criterion of ψ-optimality was derived earlier under the assumptions that, conditionally on τ, the observation errors $\mathbf{e}$ have a normal distribution and that θ has a normal prior distribution. It is shown in Pilz (1981a) that ψ-optimality is also appropriate under a variety of relaxations of these assumptions, extending the results of Rao (1976). Pilz considers situations where the prior distribution of θ is such that $(E(\theta) = \theta_0$, $\mathrm{var}(\theta \mid \tau) = (\tau R)^{-1}$ and $E(\tau^{-1}) < \infty$ and the observation errors $\mathbf{e}$ are such that $E(\mathbf{e}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{e}) = \tau^{-1}I$.

With $\theta_1 = (R + XX^T)^{-1}(X\mathbf{y} + R\theta_0)$ Pilz shows that $\mathbf{c}^T\theta_1$ is a Bayes estimator of $\mathbf{c}^T\theta$ among all linear estimators under squared error loss. The corresponding expected risk is proportional to $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$. Furthermore he shows that $\mathbf{c}^T\theta_1$ is minimax-Bayes over all estimators with the maximum expected risk again being proportional to $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$. We may note, however, that it is demonstrated in Goel and DeGroot (1980) that under mild regularity conditions if the posterior mean of θ is θ_1, a linear estimator, then conditionally on τ, the prior distribution of θ and the distribution of the errors $\mathbf{e}$ must be normal. Hence, ψ-optimality may only be appropriate in a strictly Bayesian sense under the assumptions of normality and squared error loss.

Bayesian experimental design is discussed in Sinha (1970) and Owen (1970). Owen considers design for analysis of variance models as do Giovagnoli and Verdinelli (1982) and Verdinelli (1982). Bandemer and Pilz (1978) and Pilz (1979a, b, c, d) also discuss ψ-optimality and discuss in detail the case where the matrix ψ is of full rank. Extensions of this work appear in Pilz (1981a, b), Näther and Pilz (1980) and Gladitz and Pilz (1982a, b). Pilz and his co-workers have been primarily responsible for developing the mathematical properties and applicability of Bayesian optimal designs. Some of their results will be discussed in the next section. Brooks (1972, 1974, 1976) uses the ψ-optimality criterion to derive optimal designs for prediction and also considers the choice of which independent variables to include in the model. Brooks (1977) also discusses the problem of designing an experiment to be able to control the response at a particular value.

An alternative to ψ-optimality in a Bayesian context is the equivalent of D-

optimality—that is to maximize the determinant of $R + XX^T$. This is discussed in Stone (1959), Sinha (1970), Guttman (1971) and Smith and Verdinelli (1980). This approach will maximize the increase in Shannon information in the experiment (Lindley, 1956). Chernoff (1972, page 37) expresses some criticisms of D-optimality in that using this criterion may be mathematically appealing but tends to avoid the issue of specifying what use is to be made of the experiment and what loss function is appropriate. In the context of augmenting previously chosen designs we find papers discussing the maximization of the determinant of $(X_0X_0^T + XX^T)$ where $X_0X_0^T$ is fixed. This is discussed in Covey-Crump and Silvey (1970), Dykstra (1971), Mayer and Hendrickson (1973), Evans (1979) and Johnson and Nachtsheim (1983).

In Section 2 we will discuss the general equivalence theorem of optimal design as applied to ψ-optimality. A bound on the minimum number of points in a ψ-optimal design is found which is an improvement on bounds given in the previous literature. The special case of **c**-optimality is investigated in Section 3. A new geometrical interpretation of Bayesian **c**-optimal design is given and a parallel of Elfving's Theorem is derived. The approximation to integer designs is discussed in Section 4 and it is shown how rounding noninteger designs to integer designs leads to designs which are almost optimal.

Some examples for particular design spaces are given in Section 5. Finally, in Section 6, the important special case of polynomial regression is considered and the geometrical results of Section 3 are used to find optimal designs.

2. Equivalence theorem. A ψ-optimal design leads to the matrix XX^T minimizing $\operatorname{tr} \psi(R + XX^T)^{-1}$, with $XX^T = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$ for $\mathbf{x}_i \in \mathscr{X}$, $i = 1, \cdots, n$. Rather than consider this integer programming problem directly we will follow usual practice and introduce a continuous relaxation. Define the set H as the set of all probability measures on the design space X and define

$$\mathscr{S} = \{(R + XX^T) \mid XX^T = n \int_{\mathscr{X}} \mathbf{x}\mathbf{x}^T \, d\zeta(\mathbf{x}),\ \zeta \in H\}.$$

We define the function ϕ on $\mathscr{S}$ by $\phi(M) = \operatorname{tr} \psi M^{-1}$ and minimize ϕ over $\mathscr{S}$. The set $\mathscr{S}$ is convex. It is the closed convex hull of $\{R + n\mathbf{x}\mathbf{x}^T \mid \mathbf{x} \in \mathscr{X}\}$, the set of possible $R + XX^T$ matrices obtained by one-point designs. The convexity of $\mathscr{S}$ ensures that there always exists a discrete measure on $\mathscr{X}$ solving the approximate problem. That is, there exists an integer m, $m \leq k(k + 1)/2 + 1$, such that the optimal XX^T can be written $XX^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$, $\mathbf{x}_i \in \mathscr{X}$, $n_i > 0$ and $\sum_{i=1}^m n_i = n$. Note that for Bayesian optimal designs in general the optimal measure on $\mathscr{X}$ will depend on n, the number of observations to be taken. This is not the usual case in classical design theory. Hence, rather than refer to optimal probability measures on $\mathscr{X}$ we will refer to the optimal approximate design matrix XX^T in full as the optimal design.

The notation that observations may be taken at points $\mathbf{x} \in \mathscr{X}$ is taken to include arbitrary linear regression models. For example, regressions with an intercept term may be such that the first coordinate of $\mathbf{x} \in \mathscr{X}$ is constrained to

be one, in a polynomial regression of degree p we have $\mathbf{x} = (1, x, x^2, \cdots, x^p)^T$, or, in general we may have $\mathbf{x} = (f_1(\mathbf{z}), f_2(\mathbf{z}), \cdots, f_k(\mathbf{z}))$ where $\mathbf{z} \in \mathscr{Z}$ for some suitable set $\mathscr{Z}$.

Note that taking an observation at $\mathbf{x}$ is exactly equivalent to taking one at $-\mathbf{x}$. So henceforth for mathematical convenience we will assume the set $\mathscr{X}$ to be symmetric. Thus we extend $\mathscr{X}$ to include not only points $\mathbf{x}$ at which observations may be taken but also all corresponding points $-\mathbf{x}$. For example suppose we have a quadratic regression model with the independent variable x constrained to lie in the interval $[a, b]$ then the set $\mathscr{X}$ consists of all points $\mathbf{x} = (1, x, x^2)^T$ and $\mathbf{x} = (-1, -x, -x^2)^T$ for $x \in [a, b]$. In general we will assume $\mathscr{X}$ is closed, bounded and symmetric. The set $\mathscr{X}$ may be a finite collection of points or a closed and bounded region.

The function $\phi(\cdot)$ is convex on $\mathscr{S}$ and is continuous and differentiable everywhere on $\mathscr{S}$. If rank $(\psi) = k$, $\phi(\cdot)$ is strictly convex and there is a unique minimum. If rank $(\psi) < k$, $\phi(\cdot)$ is convex but not strictly convex, and there may be a convex subset of $\mathscr{S}$ at which the minimum occurs. In either case define the Fréchet directional derivative at $M_0 = R + X_0X_0^T$ in the direction of $M_1 = R + X_1X_1^T$ as:

$$F_\phi(M_0, M_1) = \lim_{\varepsilon \downarrow 0}[\phi\{(1 - \varepsilon)M_0 + \varepsilon M_1\} - \phi\{M_0\}].$$

Using the following matrix identity for any $\varepsilon > 0$

$$\begin{aligned}(M_0 + \varepsilon(M_1 - M_0))^{-1} &= M_0^{-1} - \varepsilon M_0^{-1}(M_1 - M_0)M_0^{-1} \\ &\quad + \varepsilon^2 M_0^{-1}(M_1 - M_0)(M_0 + \varepsilon(M_1 - M_0))^{-1}(M_1 - M_0)M_0^{-1}\end{aligned} \tag{2.1}$$

we see that

$$F_\phi(M_0, M_1) = \operatorname{tr}[\psi M_0^{-1}(M_0 - M_1)M_0^{-1}]. \tag{2.2}$$

Note that (2.2) may be written:

$$F_\phi(M_0, M_1) = \operatorname{tr}[\psi M_0^{-1}(X_0X_0^T - X_1X_1^T)M_0^{-1}].$$

We may now apply the general equivalence theorem of Whittle (1973) as given in Silvey (1980).

THEOREM 1. *Any of the following* 3 *conditions are necessary and sufficient for* $X_0X_0^T$ *to be an optimal design:*

(i) $F_\phi(R + X_0X_0^T, R + X_1X_1^T) \geq 0$ *for all* $R + X_1X_1^T \in \mathscr{S}$
(ii) $F_\phi(R + X_0X_0^T, R + n\mathbf{x}\mathbf{x}^T) \geq 0$ *for all* $\mathbf{x} \in \mathscr{X}$
(iii) $\min_{x \in \mathscr{X}} F_\phi(R + X_0X_0^T, R + n\mathbf{x}\mathbf{x}^T)$
$= \max_{R + XX^T \in \mathscr{S}} \min_{\mathbf{x} \in \mathscr{X}} F_\phi(R + XX^T, R + n\mathbf{x}\mathbf{x}^T)$.

Furthermore, if $X_0X_0^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$ *where* $\mathbf{x}_i \in \mathscr{X}$, $n_i > 0$ *and* $\sum_{i=1}^m n_i = n$, *then for* $i = 1, \cdots, m$

$$F_\phi(R + X_0X_0^T, R + n\mathbf{x}_i\mathbf{x}_i^T) = 0.$$

The proof of Theorem 1 directly parallels the proofs in Silvey (1980) page 19–

23 for general linear criterion functions. This theorem is also given in a slightly different form in Bandemer and Pilz (1978).

It is interesting to note that condition (ii) reduces to the condition that if $X_0X_0^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$, $\mathbf{x}_i \in \mathscr{X}$, $n_i > 0$, $\sum_{i=1}^m n_i = n$ then for all $\mathbf{x}_i = 1, \cdots, m$ and all $\mathbf{y} \in \mathscr{X}$

$$\mathbf{x}_i^T(R + X_0X_0^T)^{-1}\psi(R + X_0X_0^T)^{-1}\mathbf{x}_i \geq \mathbf{y}^T(R + X_0X_0^T)^{-1}\psi(R + X_0X_0^T)^{-1}\mathbf{y}.$$

Thus, the points at which to take observations lie on a surface $\mathbf{x}^TA\mathbf{x} = \lambda > 0$, centered at the origin, containing the set $\mathscr{X}$ and intersecting $\mathscr{X}$ on the boundary of $\mathscr{X}$. A ψ-optimal design will only include points which lie in the intersection of this surface and the boundary of $\mathscr{X}$. Note also that if $\text{rank}(\psi) = k$ the surface is an ellipsoid containing $\mathscr{X}$.

We will now use Theorem 1 to derive an upper bound on the minimum number of points at which a ψ-optimal design must take observations. From the convexity of $\mathscr{S}$ we already have a bound of $k(k + 1)/2 + 1$, and we proceed to improve on this bound in Theorem 2. The ideas of Theorem 2 parallel those of Chernoff (1953) who derived a similar result for designs minimizing the average asymptotic variance of certain estimators in models that are not necessarily linear.

THEOREM 2. *Let* $\text{rank}(\psi) = r$. *There is a* ψ*-optimal design minimizing* $\text{tr}\,\psi(R + XX^T)^{-1}$ *that includes at most* $r(2k - r + 1)/2$ *different values of* $\mathbf{x}_i \in \mathscr{X}$.

PROOF. Using a linear transformation, we assume without loss of generality that $\psi_{ij} = 0$, $i \neq j$; $\psi_{ii} = 1$, $i = 1, \cdots, r$; and $\psi_{ii} = 0$, $i = r + 1, \cdots, k$. We consider $\mathscr{S}$ to be a subset of $k(k + 1)/2$ dimensional Euclidean space. Let M_0 be an element of $\mathscr{S}$ where $\text{tr}\,\psi M^{-1}$ is minimized. By constructing a $(k(k + 1)/2) - 1$ dimensional affine set H_1 and showing that H_1 is a supporting hyperplane to $\mathscr{S}$ at M_0 we shall prove that M_0 is a boundary point of $\mathscr{S}$. As M_0 is optimal we have for all $M_1 \in \mathscr{S}$

$$F_\phi(M_0, M_1) \geq 0$$

or equivalently,

$$\sum_{i=1}^r [M_0^{-1} - M_0^{-1}M_1M_0^{-1}]_{ii} \geq 0.$$

Define H_1 by the linear constraint

$$\sum_{i=1}^r [M_0^{-1} - M_0^{-1}MM_0^{-1}]_{ii} = 0.$$

Clearly M_0 lies in H_1 which is a supporting hyperplane to $\mathscr{S}$. The set $H_1 \cap \mathscr{S}$ is a closed convex set on the boundary of $\mathscr{S}$ with extreme points being extreme points of $\mathscr{S}$. If $r = k$ the theorem is proved. For $r < k$ consider the $k(k + 1)/2 - r(2k - r + 1)/2$ dimensional affine set H_2 defined by the $r(2k - r + 1)/2$ linear constraints

$$[M_0^{-1}M - I]_{ij} = 0 \quad i = 1, 2, \cdots, r, j = i, i + 1, \cdots, k.$$

Clearly $H_2 \subset H_1$ and all points in $\mathscr{S} \cap H_2$ correspond to optimal designs. The set $\mathscr{S} \cap H_2$ is a closed convex set on the boundary of $\mathscr{S}$. Take M_2 to be an extreme point of $H_2 \cap \mathscr{S}$. The point M_2 corresponds to an optimal design and is

also a point in $H_1 \cap \mathscr{S}$. Let m be the least number of extreme points of $H_1 \cap \mathscr{S}$ required to express M_2 as a convex combination. Denote by H_3 the set of linear combinations of these points. Clearly $H_3 \subset H_1$ and $\dim(H_3) = m - 1$. Furthermore, M_2 is an interior point of $H_3 \cap \mathscr{S}$ since if it were a boundary point we would only need $m - 1$ points to generate M_2. Hence $H_3 \cap H_2 = \{M_2\}$ and $\dim(H_3 \cap H_2) = 0$ because otherwise M_2 would be expressible as a convex combination of the distinct points in $H_2 \cap \mathscr{S}$. Let $H_2 \oplus H_3$ denote the space spanned by H_2 and H_3. As H_2 and H_3 are contained in H_1 we have $H_2 \oplus H_3 \subseteq H_1$ and thus $\dim(H_2) + \dim(H_3) - \dim(H_2 \cap H_3) \leq \dim(H_1)$. Substituting into the above expression gives the inequality

$$m \leq r(2k - r + 1)/2.$$

Thus we have the desired result that an optimal design exists on at most $r(2k - r + 1)/2$ different points $\mathbf{x} \in \mathscr{X}$.

Note that when ψ is of full rank, $r = k$, the bound is $k(k + 1)/2$ and when $r = 1$, that is $\psi = \mathbf{c}\mathbf{c}^T$ for some $\mathbf{c}$, the bound is k. This bound is an improvement on $k(k + 1)/2 + 1$, the usual bound from Carathéodory's theorem. We will also see in subsequent examples and in Section 3 that we may often need less than $r(2k - r + 1)/2$ points. Note also that the proof of Theorem 2 relies on elements of $\mathscr{S}$ being nonsingular, even when XX^T is singular.

It is also interesting to note that the bound of Theorem 2 also applies if we consider designs taking observations only at the extreme points of the symmetric convex hull of $\mathscr{X}$. It is shown in Chaloner (1982) that for any design giving $X_1X_1^T$ there is a design giving $X_2X_2^T$ taking observations only at points in $\mathscr{X}$ which are extreme points of the convex hull of $\mathscr{X}$ such that $\operatorname{tr} \psi(R + X_2X_2^T)^{-1} \leq \operatorname{tr} \psi(R + X_1X_1^T)^{-1}$ for any ψ. This parallels the result of Ehrenfeld (1959) who showed that the class of designs supported by the extreme points of $\mathscr{X}$ is an essentially complete class. Thus an optimal design need only include points which are extreme points of the symmetric convex hull of $\mathscr{X}$. Instead of minimizing $\phi(\cdot)$ on the set $\mathscr{S}$ we may consider minimizing $\phi(\cdot)$ on the subset of $\mathscr{S}$ corresponding to designs on these extreme points. The proof of Theorem 2 can be adapted directly to show there is an optimal design involving at most $r(2k - r + 1)/2$ extreme points.

It has been assumed that an optimal design, taking n_i observations at $\mathbf{x}_i$, need not be an integer design. The result of Theorem 2 does not give a bound on the number of design points for an optimal integer design. The procedure of rounding noninteger n_i's to integers is discussed in Section 4.

3. c-optimality. An important special case of ψ-optimality is the situation where $\operatorname{rank}(\psi) = 1$, i.e., $\psi = \mathbf{c}\mathbf{c}^T$ for some k-dimensional vector $\mathbf{c}$. A ψ-optimal design will be that for which $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$ is minimized. That is we minimize the posterior variance of a linear combination of the parameters θ or we minimize the predictive variance at a particular point. This criterion is often referred to as $\mathbf{c}$-optimality.

Elfving (1952) gave a geometric interpretation for $\mathbf{c}$-optimal designs in classical linear regression. He showed that if $\mathscr{X}$ is symmetric, closed and convex then a

design minimizing the variance of the least squares estimate of $\mathbf{c}^T\theta$ need take observations at only one point. Theorem 3 gives a geometric interpretation for Bayesian **c**-optimal designs and the corollary gives a similar result to Elfving's for the optimality of one-point designs.

One-point designs may not be appealing to the applied statistician in general as no information is provided on the adequacy of the model. When prior knowledge is available, however, perhaps in the form of previous data which support the model assumed, the experimenter may perhaps be willing to use a one-point design. It would always seem sensible to compute the expected loss associated with an optimal one-point design in order to have a benchmark against which to compare other designs.

Throughout this section it will be assumed that the set $\mathscr{X}$ is symmetric. Again, if the set of points $\mathbf{x}$ at which observations may be taken is not symmetric we lose no generality in including all points $-\mathbf{x}$ and making the set $\mathscr{X}$ symmetric.

THEOREM 3. *A design* $X_0X_0^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$ *is optimal for minimizing* $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$ *if and only if* $(R + X_0X_0^T)^{-1}\mathbf{c}$ *is normal to a supporting hyperplane of the convex hull of* $\mathscr{X}$ *at* $\mathbf{x}_i$ *and* $-\mathbf{x}_i$ *for* $i = 1, \cdots, m$.

PROOF. Since $\mathscr{X}$ is symmetric, a normal to a supporting hyperplane at a point $\mathbf{y}$ on the boundary of the convex hull of $\mathscr{X}$ is also a normal to a supporting hyperplane at $-\mathbf{y}$. Furthermore a vector $\mathbf{p}$ is normal to supporting hyperplanes at $\mathbf{y}$ and $-\mathbf{y}$ if and only if $(\mathbf{y}^T\mathbf{p})^2 \geq (\mathbf{x}^T\mathbf{p})^2$ for all $\mathbf{x} \in \mathscr{X}$.

Using condition (ii) of Theorem 1 we see that $X_0X_0^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$ is ψ-optimal if and only if for all $\mathbf{x} \in \mathscr{X}$

$$\sum_{i=1}^m n_i(\mathbf{x}_i^T(R + XX^T)^{-1}\mathbf{c})^2 \geq n(\mathbf{x}^T(R + X_0X_0^T)^{-1}\mathbf{c})^2.$$

This condition leads to the required result. It also follows that in a **c**-optimal design, all $\mathbf{x}_i$ must lie in the same supporting hyperplane of the convex hull of $\mathscr{X}$ or the supporting hyperplane symmetric to it.

COROLLARY. *If* $\mathscr{X}$ *is convex then a* **c***-optimal design can be concentrated at a single point* $\mathbf{x} \in \mathscr{X}$.

PROOF. Suppose that an optimal design is $X_0X_0^T = \sum_{i=1}^m n_i\mathbf{x}_i\mathbf{x}_i^T$, and denote $M_0 = R + X_0X_0^T$. As $\mathscr{X}$ is symmetric, we can without loss of generality choose the $\mathbf{x}_i$ such that $\mathbf{x}_i^TM_0^{-1}\mathbf{c} = k > 0$. Let $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^m n_i\mathbf{x}_i$, then the point $\bar{\mathbf{x}}$ lies in $\mathscr{X}$ on a supporting hyperplane through $\mathbf{x}_i$ $i = 1, \cdots, m$ with normal $M_0^{-1}\mathbf{c}$. Consider the one point design taking n observations at $\bar{\mathbf{x}}$ and denote $M_1 = R + n\,\bar{\mathbf{x}}\bar{\mathbf{x}}^T$. Using identity (2.1) with $\varepsilon = 1$ and noting that $(M_1 - M_0)\mathbf{c} = \mathbf{0}$, the $k \times 1$ vector with zero entries, we see that

$$\mathbf{c}^TM_0^{-1}\mathbf{c} = \mathbf{c}^TM_1^{-1}\mathbf{c}.$$

Thus the one-point design concentrated at $\bar{\mathbf{x}}$ is **c**-optimal.

We may use Theorem 3 to find **c**-optimal designs even when $\mathscr{X}$ is not convex.

We may suppose we can take observations anywhere in the symmetric convex hull of $\mathscr{X}$ and find the optimal one-point design. That is we find $\bar{\mathbf{x}}$ such that $(R + n\,\bar{\mathbf{x}}\bar{\mathbf{x}}^T)^{-1}\mathbf{c}$ is normal to a supporting hyperplane at $\bar{\mathbf{x}}$. Then we find the extreme points $\mathbf{x}_i$, $i = 1, \cdots, k$ such that $\bar{\mathbf{x}} = \sum_{i=1}^{k} \alpha_i \mathbf{x}_i$. Either $\mathbf{x}_i$ or $-\mathbf{x}_i$ is a point at which observations may be taken and an optimal design is to take $n\alpha_i$ observations at $\mathbf{x}_i$, $i = 1, \cdots, k$.

Note that for a regression model with an intercept term the set $\mathscr{X}$ is not convex and the corrollary does not necessarily apply. In simple linear regression, for example, if we could take observations at points $\mathbf{x} = (1, x)^T$ where $-1 \le x \le 1$ then $\mathscr{X} = \{(\pm 1, x) \mid -1 \le x \le 1\}$ which is not convex.

It is interesting to note that, for small values of n or a very precise prior distribution, one-point designs may be optimal even when $\text{rank}(\psi) > 1$. An example is given in Bandemer and Pilz (1978) and these designs are discussed further in Pilz (1981b).

4. The approximation to integer designs. Throughout this paper optimal designs are found using a continuous relaxation of the integer optimization problem. That is, we consider designs taking n_i observations at the points $\mathbf{x}_i$ where the n_i's can take nonnegative noninteger values. Of course, a true optimal design must have integer values for the n_i's, and the unconstrained minimum of $\text{tr}\,\psi(R + XX^T)^{-1}$ is not usually attained on integer values. A common practice is to round the noninteger n_i's to integers n_i^* in some systematic way; see, e.g., Fedorov (1972). Then the design taking n_i^* observations at the same points $\mathbf{x}_i$ is used. If this procedure is followed, a simple expression can be derived for the increase in the value of $\text{tr}\,\psi(R + XX^T)^{-1}$.

Let the noninteger optimal design be $X_0X_0^T = \sum_{i=1}^{m} n_i\mathbf{x}_i\mathbf{x}_i^T$ and let $M_0 = R + X_0X_0^T$. Let the integer alternative on the same $\mathbf{x}_i$'s be $X_1X_1^T = \sum_{i=1}^{m} n_i^*\mathbf{x}_i\mathbf{x}_i^T$, with $\sum_{i=1}^{m} n_i = \sum_{i=1}^{m} n_i^* = n$, and let $M_1 = R + X_1X_1^T$. From expression (2.1) with $\varepsilon = 1$, we have:

$$\begin{aligned}\text{tr}\,\psi M^{-1} &= \text{tr}\,\psi M_0^{-1} - F_\phi(M_0, M_1) \\ &\quad + \text{tr}[\psi M_0^{-1}(X_1X_1^T - X_0X_0^T)M_1^{-1}(X_1X_1^T - X_0X_0^T)M_0^{-1}].\end{aligned} \tag{4.1}$$

If we write

$$X_1X_1^T - X_0X_0^T = \sum_{i=1}^{m} (n_i^* - n_i)\mathbf{x}_i\mathbf{x}_i^T$$

and use part (ii) of Theorem 1 we see that $F_\phi(M_0, M_1) = 0$. Hence

$$\begin{aligned}&\text{tr}\,\psi M_1^{-1} - \text{tr}\,\psi M_0^{-1} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} (n_i^* - n_i)(n_j^* - n_j)[\text{tr}(\psi M_0^{-1}\,\mathbf{x}_i\mathbf{x}_i^T M_1^{-1} x_j x_j^T M_0^{-1})].\end{aligned} \tag{4.2}$$

Clearly, if n_i^* is close to n_i for all i the difference (4.2) is very small. In Theorem 5 a bound is found for this difference which, for appropriate choices of the n_i^*'s is $O(n^{-3})$. A proof of Theorem 4 is given in Chaloner (1982).

THEOREM 4. *Let* $\text{rank}(\psi) = r$, $1 \le r \le k$, *and denote a* ψ*-optimal design for* n *observations as* $X_0X_0^T$ *taking* n_i *observations at* $\mathbf{x}_i$ *for* $i = 1, 2, \cdots, m$ *and* $n_i > 0$.

Let an integer design, denoted $X_1X_1^T$ *take* n_i^* *observations at* $\mathbf{x}_i$ $i = 1, 2, \cdots, m$ *where* $\sum_{i=1}^m n_i^* = n$ *and* n_i^* *is an integer. Denote* $M_0 = R + X_0X_0^T$ *and* $M_1 = R + X_1X_1^T$, *then*

$$\frac{\text{tr } \psi M_1^{-1}}{\text{tr } \psi M_0^{-1}} \leq 1 + \frac{(\sum_{i=1}^m | n_i - n_i^* |)^2}{n(\min_i n_i^*)}.$$

Note that if we consider designs for each value of n with $\sum_{i=1}^m | n_i - n_i^* |$ uniformly bounded then in situations where $(\min_i n_i)^{-1} = O(n^{-1})$ then we also have $(\min_i n_i^*)^{-1} = O(n^{-1})$ and hence

$$\frac{\text{tr } \psi M_1^{-1}}{\text{tr } \psi M_0^{-1}} = 1 + O(n^{-2})$$

and tr ψM_1^{-1} − tr $\psi M_0^{-1} = O(n^{-3})$. This is parallel to a result given in Kiefer (1971) for classical designs. Theorem 2 indicates that there is an optimal design for each value of n with $m \leq r(2k - r + 1)/2$. Thus the condition that $\sum_{i=1}^m | n_i - n_i^* |$ is uniformly bounded is not unreasonable. We may also note here a major difference between Bayesian optimal designs and classical optimal designs. Classical optimal designs can be thought of as measures on $\mathscr{X}$ which do not depend on n. For Bayesian ψ-optimal designs m, $\mathbf{x}_i$ and n_i depend on n.

In finding an optimal integer design we must, of course, consider the possibility of taking observations at values of $\mathbf{x}$ other than those in the noninteger optimal design. However, Theorem 4 indicates that we may lose little, in terms of the expected loss, by using the same values of $\mathbf{x}_i$ and rounding the n_i's to integers. Note that the true optimal integer design has a loss associated with it which lies between tr ψM_0^{-1} and tr ψM_1^{-1}.

5. Some examples. The results of the previous sections will now be used in finding ψ-optimal designs for particular sets $\mathscr{X}$. We first consider a spherical $\mathscr{X}$ of radius b: that is

$$\mathscr{X} = \{\mathbf{x} \mid \mathbf{x}^T\mathbf{x} \leq b^2\}. \tag{5.1}$$

This would correspond to a linear regression model without a constant term. Suppose that rank$(\psi) = k$ and let $\psi^{1/2}$ denote the symmetric square root of ψ. Define

$$\lambda = (nb^2 + \text{tr } R)(\text{tr } \psi^{1/2})^{-1}$$

and let

$$X_0X_0^T = \lambda\psi^{1/2} - R.$$

If n or b^2 is large enough then $X_0X_0^T$ is positive semidefinite. Let $\mathbf{e}_i$, $i = 1, \cdots, k$ be the unit eigenvectors of $X_0X_0^T$ and μ_i be the corresponding eigenvalues. We obtain the design matrix $X_0X_0^T$ by taking observations at the points $b\mathbf{e}_i$ on the surface of the sphere and taking n_i proportional to μ_i, $i = 1, \cdots, k$. Let $M_0 = R + X_0X_0^T$ and so for any $M_1 = R + X_1X_1^T \in \mathscr{S}$ we have

$$F_\phi(M_0, M_1) = \lambda^{-2}\text{tr}(X_0X_0^T - XX^T) = \lambda^{-2}(nb^2 - \text{tr } XX^T).$$

But tr $XX^T \leq nb^2$ with equality if and only if all observations are taken on the surface of the sphere. Thus for all $M_1 \in \mathscr{S}$, $F_\phi(M_0, M_1) \geq 0$ and $X_0X_0^T$ is ψ-optimal. The optimality of $X_0X_0^T$ can also be shown explicitly by using Lagrange multipliers and differentiation (Chaloner, 1982).

Now consider the situation where rank(ψ) = 1, that is $\psi = \mathbf{c}\mathbf{c}^T$ for some k-dimensional vector $\mathbf{c}$. For a spherical $\mathscr{X}$ as in (5.1) the corollary to Theorem 3 indicates that there is an optimal one-point design. The normal at $\mathbf{x}$, on the surface of $\mathscr{X}$, is parallel to $\mathbf{x}$. Thus it is easy to show that the optimal design takes all observations at $\mathbf{x}_0$ where $\mathbf{x}_0$ is on the surface of $\mathscr{X}$ and $\mathbf{x}_0$ is parallel to $(I + (nb^2)^{-1}R)^{-1}\mathbf{c}$. This is a generalization of a result by Silvey (1969) who considered the case $n = 1$ in the context of augmenting a previous design.

As a numerical example, consider the case when $R = (r_{ij})_{ij=1}^3$ and $r_{ii} = 1$, $r_{ij} = -.25$ $i \neq j$. This would arise for example if the θ_i's were believed to be exchangeable. Furthermore suppose that ψ is diagonal with $\psi_{11} = 1$ and $\psi_{22} = \psi_{33} = 4$. This would correspond to estimating θ_1, $2\theta_2$ and $2\theta_3$ with equal weight. If $b = 1$ and $n = 12$ the optimal design is to take observations at $\mathbf{x}_1 = (.101, .703, .703)^T$, $\mathbf{x}_2 = (0, -.707, .707)^T$ and $\mathbf{x}_3 = (.994, -.076, -.076)^T$ with $n_1 = 5.29$, $n_2 = 4.75$ and $n_3 = 1.96$. We could round this noninteger design to an integer design, for example, $n_1^* = 5$, $n_2^* = 5$ and $n_3^* = 2$. It is interesting to note that the optimal design has an expected loss of $5/3 = 1.667$ and the integer design an expected loss of 1.670. Thus the integer design has an associated risk which is almost optimal.

Consider now another example for the same matrix R and the same space $\mathscr{X}$ but with $n = 1$ and $\psi = \mathbf{c}\mathbf{c}^T$ where $\mathbf{c} = (1, 0, 0)^T$. That is we have one observation and we wish to estimate θ_1. The optimal choice of $\mathbf{x}$ is $\mathbf{x}_0 = (.98, .14, .14)^T$. The classical approach, with no prior information, would give a $\mathbf{c}$-optimal design at $\mathbf{x}_0 = (1, 0, 0)^T$.

We have considered the case where $\mathscr{X}$ is spherical, that is the model has no intercept term. The case where the model has a constant term is more complicated analytically. Suppose we have a design space $\mathscr{X}$ such that for $\mathbf{x} = (x_1, \cdots, x_k)^T \in \mathscr{X}$, $x_1 = 1$ and $\sum_{i=2}^k x_i^2 \leq b^2$. Brooks (1976) solved the optimal design problem for $\psi = I$ for this space $\mathscr{X}$. His proof can be adapted for any matrix ψ of full rank (Chaloner, 1982). Gladitz and Pilz (1982b) consider ψ-optimality for a spherical space $\mathscr{X}$ and for rank(ψ) = k and give optimal noninteger and integer designs explicitly.

We will now consider the case where the space $\mathscr{X}$ is rectangular. We begin by showing how designs for a model with a constant term can be derived from those for models without a constant term.

First, suppose that we have a model without a constant term and the space $\mathscr{X}$ is a rectangle, symmetric with respect to the origin. This would arise if the constraints on $\mathbf{x}$ were of the form $-a_i \leq x_i \leq a_i$ for $i = 1, 2, \cdots, k$, for each coordinate of $\mathbf{x}$. With a linear transformation of the $\mathbf{x}$'s and of the parameters of the model, $\mathscr{X}$ can be transformed to a cube $\mathscr{C}$ centered at the origin with each side of length two units. The symmetric convex hull of $\mathscr{C}$ is just $\mathscr{C}$.

Second, suppose we have a model with a constant term and constraints on the coordinates of $\mathbf{x}$ of the form $x_1 = 1$ and $a_i \leq x_i \leq b_i$ for $i = 2, \cdots, k$. With a linear

transformation on $\mathscr{X}$, namely $z_1 = x_1$, $z_i = (2x_i - a_i - b_i)/(b_i - a_i)$ for $i = 2, \cdots, k$, we can transform $\mathscr{X}$ to a space $\mathscr{X}^*$ with

$$\mathscr{X}^* = \{\mathbf{z} \mid z_1 = \pm 1, -1 \leq z_i \leq 1 \quad \text{for} \quad i = 2, \cdots, k\}.$$

Thus the symmetric convex hull of $\mathscr{X}^*$ is again a cube $\mathscr{C}$ centered at the origin of side length two units.

In either case we can consider, without loss of generality, the cube $\mathscr{C}$ centered at the origin. It will be convenient to do so. There are 2^k extreme points of $\mathscr{C}$, i.e. the corners, so we only need to consider designs taking observations at these points. Furthermore, there are only 2^{k-1} different values of $\mathbf{x}_i\mathbf{x}_i^T$ where $\mathbf{x}_i$ is a corner. Therefore, without loss of generality we can consider taking observations at the corners where the first coordinate is one. This will give $(XX^T)_{ii} = n$, $i = 1, \cdots, k$ for designs on the corners of $\mathscr{X}$.

Suppose we can find an $X_0X_0^T$ with $(X_0X_0^T)_{ii} = n$, $i = 1, \cdots, k$ and for some diagonal matrix Λ with nonnegative entries then $M_0 = R + X_0X_0^T$ is such that $M_0 \Lambda M_0 = \psi$. It is easy to verify that $F_\phi(M_0, M_1) \geq 0$ for all $M_1 \in \mathscr{S}$. Thus, if there is such an $M_0 \in \mathscr{S}$ it is ψ-optimal. Whereas there always exists a positive semidefinite matrix M_0 satisfying $M_0 \Lambda M_0 = \psi$ and $(M_0)_{ii} = r_{ii} + n$, it need not necessarily be an element of $\mathscr{S}$. If ψ is of full rank, however, and n is large enough there will exist such a solution in $\mathscr{S}$.

As a numerical example suppose we have $k = 3$, a regression in two variables with an intercept term and the symmetric convex hull of $\mathscr{X}$ is $\mathscr{C}$. Suppose the measure μ is uniform over the face of $\mathscr{C}$, $\{\mathbf{c} \mid c_1 = 1, -1 \leq c_2, c_3 \leq 1\}$. That is, we are interested in prediction at the points where the expected value of y is $\theta_1 + c_2\theta_2 + c_3\theta_3$ and $-1 \leq c_i \leq 1$ for $i = 2, 3$. This gives a diagonal matrix ψ with $\psi_{11} = 1$ $\psi_{22} = \psi_{33} = \frac{1}{3}$. Suppose that the prior information corresponds to θ_1 being independent of both slope coefficients but the slope coefficients are positively correlated. For example, $r_{ii} = 3$, $i = 1, 2, 3$, $r_{12} = r_{13} = 0$ and $r_{23} = -2$. Then the optimal design taking observations at the corners of $\mathscr{C}$ with $M_0 \Lambda M_0 = \psi$ is given by taking observations at $\mathbf{x}_1 = (1, 1, 1)^T$, $\mathbf{x}_2 = (1, -1, -1)^T$, $\mathbf{x}_3 = (1, -1, 1)^T$ and $\mathbf{x}_4 = (1, 1, -1)^T$. If we take n_i observations at $\mathbf{x}_i$ we have $n_1 = n_2 = (n + 2)/4$ and $n_3 = n_4 = (n - 2)/4$. If $(n - 2)$ is a multiple of 4 this design is an integer design and so must be the optimal integer design.

It is interesting to note that for this precision matrix R and for any diagonal matrix ψ we have this same optimal design. This is due to the fact that if we find an $M \in \mathscr{S}$ such that $M \Lambda M = \psi$ for any diagonal ψ we have the same solution. The minimum expected loss is different however, for different matrices ψ and the optimal integer design will also depend on the exact value of ψ.

Gladitz and Pilz (1982a) give an algorithm for the construction of ψ-optimal designs for general experimental regions for special combinations of R and ψ.

6. Polynomial regression. An important special case of the linear regression model is polynomial regression. Designs for classical polynomial regression models are discussed extensively in the literature, for example in Kiefer and Wolfowitz (1959), Hoel and Levine (1964), Hoel (1966, 1981), Herzberg and Cox (1972) and Studden (1968, 1971). The only reference to Bayesian optimal design

for polynomial regression is Smith and Verdinelli (1980) who use the D-optimality criterion.

It will be assumed that we have a polynomial regression of degree p and we can take observations anywhere in a closed interval. Without loss of generality we will assume that the interval is $[-1, 1]$. Thus the set of points $\mathbf{x}$ at which observations may be taken are $\{\mathbf{x} \mid \mathbf{x} = (1, x, x^2, \cdots, x^p)^T, -1 \leq x \leq 1\}$. The set $\mathscr{X}$, in the notation of previous sections, consists of these points, $\mathbf{x}$, and all points $-\mathbf{x}$.

We denote $\theta = (\theta_0, \theta_1, \cdots, \theta_p)^T$ to be the vector of unknown coefficients which are assumed to have a normal prior distribution given τ with precision matrix R where $R = (r_{ij})_{ij=0}^p$ is a known positive definite matrix.

We will restrict attention to $\mathbf{c}$-optimal designs. That is, we wish to choose a design to estimate $\mathbf{c}^T\theta$ where $\mathbf{c}$ is a $(p + 1)$-dimensional vector. For prediction $\mathbf{c} = (1, a, \cdots, a^p)^T$ where a is a real number. The case $|a| > 1$ corresponds to extrapolation and the case $|a| \leq 1$ to interpolation. These two different cases lead to different kinds of optimal designs. Note that $\mathscr{X}$ is not convex so an optimal one-point design does not necessarily exist.

It will be demonstrated that for designing to extrapolate or designing to estimate θ_p, the coefficient of x^p, the optimal design involves taking observations at the $p + 1$ Chebychev points, $x_j = -\cos \pi j p^{-1}$, $j = 0, 1, \cdots, p$. The Chebychev points are the points at which the pth Chebychev polynomial is maximized or minimized on $[-1, 1]$.

We will use the following lemma which demonstrates that the vector of Chebychev coefficients is the normal to the supporting hyperplane of the symmetric convex hull of $\mathscr{X}$ at the points corresponding to Chebychev points.

LEMMA 1. *Let d_i be the coefficient of x^i in the pth Chebychev polynomial of the first kind with leading coefficient 1, $T_p(x)$. Further let $\mathbf{x}_i = (1, x_i, x_i^2, \cdots, x_i^p)^T$, where $x_i = -\cos i\pi p^{-1}$, $i = 0, 1, \cdots, p$ are the Chebychev points. Then the vector $\mathbf{d} = (d_0, d_1, \cdots, d_{p-1}, d_p)$ is normal to the supporting hyperplane of the symmetric convex hull of $\mathscr{X}$ at $\mathbf{x}_i$, $i = 0, 1, \cdots, p$.*

PROOF. Note that for all $\mathbf{x} = (1, x, x^2, \cdots, x^p)^T$, $\mathbf{x}^T\mathbf{d} = T_p(x)$ *and* $\mathbf{x}_i^T\mathbf{d} = T_p(x_i) = (-1)^{p+1-i}2^{-p+1}$, $i = 0, 1, \cdots, p$. For all $x \in [-1, 1]$ we have

$$(T_p(x_i))^2 \geq (T_p(x))^2$$

(Karlin and Studden, 1966, page 281), with equality only at $x = x_i$, $i = 0, 1, \cdots, p$. Equivalently for all $\mathbf{x} \in \mathscr{X}$, $(\mathbf{x}_i^T\mathbf{d})^2 \geq (\mathbf{x}^T\mathbf{d})^2$ and the lemma is proved.

Thus Theorem 3 implies that if we can find a design on the Chebychev points with $M_0 = R + XX^T$ and $M_0^{-1}\,\mathbf{c} = \lambda\mathbf{d}$ for some constant λ then the design is $\mathbf{c}$-optimal. For extrapolation, or estimation of θ_p we can find such a design.

First suppose that $\mathbf{c} = (1, a, \cdots, a^p)^T$ and $|a| > 1$. That is, we wish to predict at the point $x = a$ which lies outside the experimental region. Taking n_j observations at x_j, $j = 0, 1, \cdots, p$, $x_j = -\cos \pi j p^{-1}$ and setting $\lambda\mathbf{c} = M_0\mathbf{d}$ leads to the

following equations for $\ell = 0, 1, \cdots, p$:

$$\lambda a^{\ell} = \sum_{i=0}^{p} r_{\ell i} d_i + \sum_{j=0}^{p} n_j x_j^{\ell} (-1)^{p+1-j} 2^{-p+1}. \tag{6.1}$$

Solving equations (6.1) and the equation $\sum_{j=0}^{p} n_j = n$ will lead to the optimal design. For large n the solution approaches the classical solution for extrapolation at $x = a$, given in Hoel and Levine (1965). They define the Lagrange polynomials $L_j(x)$ for $j = 0, 1, \cdots, p$ by

$$L_j(x) = \frac{\Pi_{k \neq j}(x - x_k)}{\Pi_{k \neq j}(x_j - x_k)}.$$

Then their optimal design for extrapolation at $x = a$ is to take n_j observations at $x_j, j = 0, 1, \cdots, p$ with

$$n_j = \frac{n \mid L_j(a) \mid}{\sum_{i=0}^{p} \mid L_i(a) \mid}. \tag{6.2}$$

The solution (6.2) can be obtained by substituting $r_{ij} = 0$, $i, j = 0, 1 \cdots p$, in equations (6.1). As n increases the solution to equations (6.1) approach (6.2) for any fixed R. Thus for large enough n the n_j's given by equations (6.1) will be positive and correspond to the Bayesian c-optimal design. For small values of n the solution to (6.1) may give negative values for some of the n_j's. It would seem that one approach may be to round negative n_i's to zero and round down other n_i's accordingly. This may not be the optimal solution but a few examples have been investigated where the expected loss is almost optimal following this procedure.

For all illustration consider quadratic regression. The Chebychev points are $x_0 = -1$, $x_1 = 0$ and $x_2 = 1$. If we let

$$b_1 = \frac{(2a^2 + 1)r_{02} - a^2 r_{00} - 2r_{22}}{2a^2 - 1}$$

and

$$b_2 = \frac{2a^2 r_{12} + a r_{02} - 2a r_{22} - a^2 r_{01}}{2a^2 - 1}$$

then the solution given by (6.1) is

$$n_0 = \frac{na(a-1)}{2(2a^2-1)} - b_1 \frac{(a-1)}{2a} + b_2 \frac{(2a^2-1)}{2a^2}, \quad n_1 = \frac{n(a^2-1)}{2a^2-1} + b_1,$$

$$n_2 = \frac{na(a+1)}{2(2a^2-1)} - b_1 \frac{(a+1)}{2a} - b_2 \frac{(2a^2-1)}{2a^2}.$$

Note that the terms involving n correspond to the classical solution. For any given prior precision matrix R all the n_j's will be positive for n sufficiently large. For small n or very informative prior information some of the n_j's may be negative.

In a similar manner an optimal design for estimating θ_p can be found which

involves only the Chebychev points. In this case $\mathbf{c} = (0, 0, 1)^T$ and setting $\lambda\mathbf{c} = M_0\mathbf{d}$ gives:

$$\begin{aligned} &\sum_{i=1}^{p} r_{\ell i} d_i + \sum_{j=0}^{p} n_j x_j^{\ell} (-1)^{p+j-1} 2^{-p+1} = 0, \quad \ell = 0, 1, \cdots, p-1 \\ &\sum_{i=1}^{p} r_{pi} d_i + \sum_{j=0}^{p} n_j x_j^{p} (-1)^{p+j-1} = \lambda. \end{aligned} \tag{6.3}$$

Solving equation (6.3) with $\sum_{j=0}^{p} n_j = n$ gives the optimal design. Again, for large n, the solution to equations (6.3) approach the classical solution of Kiefer and Wolfowitz (1959) who showed that to estimate θ_p the optimal design takes n_j observations at the Chebychev points $x_j, j = 0, 1, \cdots, p$ with

$$n_0 = n_p = \frac{n}{2p} \quad \text{and} \quad n_j = \frac{n}{p} \quad j = 1, 2, \cdots, p-1.$$

For any prior precision matrix R equations (6.3) will lead to the optimal Bayesian design for n sufficiently large.

As an illustration consider cubic regression and designing to minimize the posterior variance of θ_3. The Chebychev points for $p = 3$ are $x_0 = -1$, $x_1 = -\frac{1}{2}$, $x_2 = \frac{1}{2}$ and $x_3 = 1$. Solving equations (6.3) gives

$$n_0 = \frac{n}{6} + \frac{4}{3}\left(\frac{3r_{11}}{4} - r_{13}\right) + \frac{48}{5}\left[\left(\frac{3r_{01}}{4} - r_{03}\right) - \left(\frac{3r_{21}}{4} - r_{23}\right)\right]$$

$$n_1 = \frac{n}{3} - \frac{4}{3}\left(\frac{3r_{11}}{4} - r_{13}\right) + \frac{8}{5}\left[\left(\frac{3r_{01}}{4} - r_{03}\right) - \left(\frac{3r_{21}}{4} - r_{23}\right)\right]$$

$$n_2 = \frac{n}{3} - \frac{4}{3}\left(\frac{3r_{11}}{4} - r_{13}\right) - \frac{8}{5}\left[\left(\frac{3r_{01}}{4} - r_{03}\right) - \left(\frac{3r_{21}}{4} - r_{23}\right)\right]$$

$$n_3 = \frac{n}{6} + \frac{4}{3}\left(\frac{3r_{11}}{4} - r_{13}\right) - \frac{48}{5}\left[\left(\frac{3r_{01}}{4} - r_{03}\right) - \left(\frac{3r_{21}}{4} - r_{23}\right)\right].$$

For interpolation in $[-1, 1]$ or estimation of an arbitrary linear combination of the coefficients, it is not necessarily optimal to take observations at the Chebychev points as simple examples will show. In a classical design to interpolate at a particular point it is optimal to take all observations at that point. With prior information however this will not necessarily be true. Taking observations at the Chebychev points and setting $M_0^{-1}\mathbf{c} = \lambda\mathbf{d}$ does not in general lead to a positive solution for the n_j's even for large n except for a few special values of R and $\mathbf{c}$. The problem of interpolation is discussed in Chaloner (1982). There appears to be no general approach for finding $\mathbf{c}$-optimal designs for this case.

The designs derived in this section for extrapolation and estimation of θ_p were derived using a different approach in Chaloner (1982). The minimization of $\mathbf{c}^T(R + XX^T)^{-1}\mathbf{c}$ was considered directly and Chebychev systems of polynomials were used. The approach was parallel to that of Kiefer and Wolfowitz (1959) for finding designs to estimate θ_p in classical polynomial regression. The geometric approach used here is much simpler and the approach appears to be unique to finding optimal Bayesian designs.

7. Discussion. Some basic properties of Bayesian ψ-optimal designs have been presented. Parallels and differences between Bayesian and classical designs have been shown. The dispersed literature on Bayesian optimal design has been reviewed.

The upper bound on the number of design points, the geometric interpretation of ψ-optimal designs and the approximate optimality of rounded designs are particular aspects of Bayesian design not previously shown in the literature. The geometric interpretation is especially useful for **c**-optimal designs and led to the designs for polynomial regression in Section 5. These designs are of special interest and it should be possible to extend these results for **c**-optimality to situations where $\text{rank}(\psi) > 1$.

Acknowledgments. I would like to thank my thesis advisor, Professor M. H. DeGroot, for his help throughout this work. I would also like to thank the referees for their valuable suggestions.

REFERENCES

BANDEMER, HANS and PILZ, JÜRGEN (1978). Optimum experimental designs for a Bayes estimator in linear regression. *Transactions of the Eighth Prague Conference.* **A** 93–102.

BROOKS, R. J. (1972). A decision theory approach to optimal regression designs. *Biometrika* **59** 563–571.

BROOKS, R. J. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika* **61** 303–311.

BROOKS, R. J. (1976). Optimal regression designs for prediction when prior knowledge is available. Metrika **23** 221–230.

BROOKS, R. J. (1977). Optimal regression design for control in linear regression. *Biometrika* **64** 319–325.

CHALONER, KATHRYN (1982). Optimal Bayesian experimental design for linear models. Ph.D. Thesis. Carnegie-Mellon Technical Report No. 238.

CHERNOFF, H. (1953). Locally optimal designs for estimating parameters. *Ann. Math. Statist.* **2** 586–602.

CHERNOFF, H. (1972). *Sequential Analysis and Optimal Design.* SIAM, Philadelphia.

COVEY-CRUMP., P. A. K. and SILVEY, S. D. (1970). Optimal regression designs with previous observations. *Biometrika* **57** 551–556.

DUNCAN, G. and DEGROOT, M. H. (1976). A mean squared error approach to optimal design theory. *Proceedings of the 1976 conference on information: sciences and systems.* The Johns Hopkins University, 217–221.

DYKSTRA, OTTO, JR. (1971). The augmentation of experimental data to maximize $|X^TX|$. *Technometrics* **13** 682–688.

EHRENFELD, S. (1956). Complete class theorems in experimental designs, in *Proceedings of the Third Berkeley Symposium*, ed. J. Neyman. University of California Press, Berkeley and Los Angeles.

ELFVING, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Statist.* **23** 255–262.

EVANS, J. W. (1979). Computer augmentation of experimental designs to maximize $|X^TX|$. *Technometrics* **21** 321–330.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments.* Trans. and ed. W. J. Studden and E. M. Klimko. Academic, New York.

GIOVAGNOLI, A. and VERDINELLI, I. (1983). Bayes *D*-optimum and *E*-optimum block designs. *Biometrika*, to appear.

GLADITZ, J. and PILZ, J. (1982a). Construction of optimal designs in random coefficient regression models. *Math. Operationsforsch. Stat.*, Ser. Statistics **13** 371–385.

GLADITZ, J. and PILZ, J. (1982). Bayes designs for multiple linear regression on the unit sphere. *Math. Operationsforsch. Stat.*, Ser. Statistics **13** 491–506.

GOEL, P. K. and DEGROOT, M. H. (1980). Only normal distributions have linear posterior expectations in linear regresson. *J. Amer. Statist. Assoc.* **75** 895–900.

GUTTMAN, IRWIN (1971). A remark on the optimal regression designs with previous observations of Covey-Crump and Silvey. *Biometrika* **50** 683–685.

HERZBERG, AGNES M. and COX, D. R. (1972). Some optimal designs for interpolation and extrapolation. *Biometrika* **59** 551–561.

HOEL, P. G. and LEVINE, A. (1964). Optimal spacing and weighting in polynomial prediction. *Ann. Math. Statist.* **35** 1553–1563.

HOEL, PAUL G. (1966). A simple solution for optimal Chebyshev regression extrapolation. *Ann. Math. Statist.* **37** 720–725.

HOEL, PAUL G. (1981). Regression systems for which optimal extrapolation designs require exactly $k + 1$ points. *Ann. Statist.* **9** 909–912.

JOHNSON, MARK E. and NACHTSHEIM, CHRISTOPHER J. (1983). Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics*, to appear.

KARLIN, S. and STUDDEN, W. J. (1966). *Tchebycheff Systems: with Applications in Analysis and Statistics.* Wiley, New York.

KIEFER, J. (1959). Optimal experimental designs, (with discussion). *J. Roy. Statist. Soc. B.* **21** 272–319.

KIEFER, J. (1961). Optimum designs in regression problems. II. *Ann. Math. Statist.* **32** 298–325.

KIEFER, J. (1971). The role of symmetry and aproximation in exact design optimality. In *Statistical Decision Theory and Related Topics*, ed. S. S. Gupta and J. Yachel. Academic, New York.

KIEFER, J. (1973). Optimum designs for fitting biased multiresponse surfaces. In *Multivariate Analysis*, Vol. 3, ed. P. R. Krishnaiah, 287–297. Academic, New York.

KIEFER, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2** 849–879.

KIEFER, J. and WOLFOWITZ, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.* **30** 271–294.

KIEFER, J. and WOLFOWITZ, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.* **14** 363–366.

KIEFER, J. and WOLFOWITZ, J. (1965). On a theorem of Hoel and Levine on extrapolation. *Ann. Math. Statist.* **36** 1627–1655.

LINDLEY, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.

LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. B* **32** 31–53.

MAYER, LAWRENCE S. and HENDRICKSON, ARLO D. (1973). A method for constructing an optimal regression design after an initial set of input values has been selected. *Comm. Statist.* **2(5)** 465–477.

NÄTHER, W. and PILZ, J. (1980). Estimation and experimental design in a linear regression model using prior information. *Zastosowania Matematyki* **16** 565–577.

OWEN, R. J. (1970). The optimum design of a two-factor experiment using prior information. *Ann. Math. Statist.* **41** 1917–1934.

PILZ, J. (1979a). Konstruktion von optimalen diskreten Versuchsplänen für eine Bayes-Schätzung im linearen Regressionsmodell. *Freiberger Forschungshefte* **117** 123–152.

PILZ, J. (1979b). Optimalitätskriterien, Zulässigkeit und Vollständigkeit im Planungsproblem für eine bayesshe Schätzung im linearen Regressionsmodell. *Freiberger Forschungshefte* **117** 67–94.

PILZ, J. (1979c). Das bayesshe Schätzproblem im linearen Regressionsmodell. *Freiberger Forschungshefte* **117** 21–55.

PILZ, J. (1979d). Entscheidungstheoretische Darstellung des problems der bayesshen Schätzung und versuchsplanung im linearen regressionsmodell. *Freiberger Forschungshefte* **117** 7–20.

PILZ, J. (1981a). Robust Bayes and minimax-Bayes estimation and design in linear regression. *Math. Operationsforsch. Stat.*, Ser. Statistics **12** 163–177.

PILZ, J. (1981b). Bayesian one-point designs in linear regression. Paper presented at the fifth international summer school on problems of model choice and parameter estimation in regression analysis, March 1981, Sellin, G.D.R.

RAO, C. R. (1976). Estimation of parameters in a linear model. *Ann. Statist.* **4** 1023–1037.

SILVEY, S. D. (1969). Multicollinearity and imprecise estimation. *J. Roy. Statist. Soc. B.* **31** 539–552.

SILVEY, S. D. (1980). *Optimal Design.* Chapman and Hall, London and New York.

SINHA, BIMAL KUMAR (1970). A Bayesian approach to optimum allocation in regression problems. *Calcutta Statistical Association Bulletin* **19** 45–52.

SMITH, A. F. M. and VERDINELLI, I. (1980). A note on Bayes designs for inference using a hierarchical linear model. *Biometrika* **67** 613–619.

STONE, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist.* **30** 55–70.

THEIL, H. (1971). *Principles of Econometrics.* Wiley, New York.

VERDINELLI, I. (1982). Computing Bayes *D*- and *A*-optimal block designs for a two-way model. Unpublished manuscript.

WHITTLE, P. (1973). Some general points in the theory of optimal experimental design. *J. Roy. Statist. Soc. B* **35** 123–130.

SCHOOL OF STATISTICS
DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108