REVIEW OF OPTIMAL BAYES DESIGNS

by

Anirban DasGupta*
Purdue University

Technical Report #95-4

Department of Statistics
Purdue University

February 1995
Revised March 1996

*

Dedicated to J.H.B. Kemperman, a unique person.

Review of Optimal Bayes Designs

by

Anirban DasGupta*
Purdue University

Abstract

This article gives a review of the optimality theory of Bayes experimental designs. The description includes common formulations, the mathematics, the existing results, and explicit examples. Since Bayes experimental designs cannot be understood separately from the rich theory of classical optimal designs, this article also gives a lot of information on the classical theory and practice. Future directions are discussed and there is an appendix on moment methods and orthogonal polynomials, which may be of independent interest. Nonlinear problems are described only tangentially in Section 11.

*Research supported by NSF grant DMS 9307727

Table of Contents

1. General Introduction.

It is fair to say that for as long as we can recall, statistical training has emphasized the role of design of an experiment in extracting the correct type of information and making accurate inferences for the problem of interest. Proper design of an experiment is evidently a crucial aspect of sound statistical practice; a classic course on design actually helps bring this out much more than a sophisticated course on the mathematics of design. Historically, design of esperiments started out with agricultural studies and therefore factorial experiments and their design aspects were the natural starting points of the theory and practice of optimum design. A little exposure to factorial experiments shows how utterly important it indeed is to use the right design to avoid the pitfalls of confounding, non-estimability, missing data, and a long list of other genuine problems. See Fisher (1949). With statistical practice changing rapidly, as an influence of advances in computer technology and the inclination among many to treat statistics as mostly data analysis, the role of design of experiments and controlled studies may get substantially diminished at a future time. However, that has not happened yet. It therefore appears appropriate that a broad overview of the history, mathematics, methods, advances,and the future of experimental design should be made. A number of such contributions already exist; some are more methodological, others more technical. Design of experiments especially is one branch of statistical theory in which frequentist and Bayesian ideas, formulations, techniques, and results go very parallel; thus, although the primary goal of this writing is to make a review of the state of the art in Bayesian design, it is quite impossible to do so in isolation of the rich history of classical optimal design. As a matter of fact, Bayes design can be understood only in the context of what is known in classical design. One reason for this is that barring a few exceptions, even the formulation of a Bayes design problem requires frequentist evaluations of a design. We will therefore have to necessarily consider classical theory and methods to some extent in this writing. Although there is bound to be some overlap between the technical contents of this chapter and others, due to the connections of the mathematics involved in various variety of optimal design theory, there are certain unique aspects of this chapter, both informationally and technically. In particular, this chapter is unique in its description of the role of the prior in determining an optimal design, how the role of the prior typically diminishes in a very strong sense with increasing

sample size, conditional optimal design formulations that would make sense only in the Bayesian context of this chapter, and sample size problems that are always regarded as important by applied scientists, especially those conducting clinical trials. On the technical side, this chapter provides a self contained brief account of two relevant but highly used tools of pure mathematics: moment methods and the algebra of orthogonal polynomials. These are presented in the appendix to retain their independent character and to allow an uninterrupted flow of the statistical content in the main body of the chapter.

In a broad intellectual sense, the problem of design is encompassing; one can make the following general statement: in any scientific problem in which the scientist has the choice and flexibility of choosing one among many initial setups of the experiment, there is an optimal design problem associated with the experiment. The concept of optimal design goes far back in the history of mathematical sciences, and probably even further back in the history of our civilization. For example, the following well known examples are all instances of optimal designs.

Example 1. Polynomial interpolation. Consider a continuous function $f(x)$ on a given bounded interval $[a, b]$. It is well known that there exists a polynomial $p(x)$ which uniformly approximates the given function $f(x)$, to any specified degree of accuracy $\epsilon > 0$, with the degree of the polynomial $p(x)$ depending on ϵ . A proof of this can be found in standard texts on analysis; for a historically important proof, see Korner (1989).

It is natural to try to find a good approximating polynomial by interpolating the function f at a fixed set of say $(n+1)$ points, $X = \{x_i, 0 \leq i \leq n\}$. Indeed, there is a unique polynomial $p(x)$ with the representation

$$p(x) \;\; = \sum_{i=0}^{n} \; l_i(x) \; f(x_i)$$

interpolating f at the points of X. In the above, $\{l_i\}$ are themselves polynomials of degree n, and are commonly known as the fundamental or cardinal polynomials of interpolation.

A natural criterion for assessing the goodness of approximation of f by p is the quantity

$$e(x) = |p(x) - f(x)|.$$

Note that $e(x)$ depends on the choice of "nodes" X. A design problem is therefore the following:

Choose a good set of nodes X according to the criterion e.

As an illustration, consider approximating the (unnormalized) Cauchy density

$$f(x) = \frac{1}{1+x^2}$$

in the interval $[-5, 5]$. Take two sets of nodes: X_1 = Equally spaced points at spacings of .5, starting at -4.75; X_2 = The points of peaks of the nth Chebyshev polynomial $T_n(x)$ in the interval $[-5, 5]$, with $n = 20$.

Straightforward computation shows that the equispaced points give a very bad fit near the boundary of the interval, and even worse, the fit deteriorates by taking more equispaced points. On the contrary, the Chebyshev nodes result in a maximum error of $< .016$ with $n = 20$. One might therefore say that X_2 is a better design than X_1 in this example. For universal results on the goodness of X_2 as the nodes of interpolation, one can see Erdos (1958) and Rivlin (1981). For further illuminating discussion of the example above, one can see Powell (1981).

Example 2. The Secretary problem. The basic Secretary problem corresponds to the situation in which n candidates are interviewed for a job in a random order and a candidate once rejected cannot be recalled. The employer would be able to rank the candidates from 1 (best) to n (worst) if s/he could indeed see them all at one time. The criterion of the employer is the following:

Maximize the probability of selecting the actual best candidate.

The following is a design problem:

What interviewing strategy should be used according to this criterion?

Although in this form, the problem is admittedly somewhat unrealistic, various modifications of the basic problem have been studied in great detail; however, even the basic problem is intellectually interesting due to the beauty and the neatness of the optimal design. The optimal design says that there exists a value $k = k(n)$ such that the employer should reject the first k candidates and then accept the very first one s/he likes better than the others who went by. In the process, there is the possibility that the employer has to

accept the last candidate to arrive, even if this is the worst candidate among all. The value k can be found with relative ease by maximizing a unimodal function defined on integers. The asymptotic solution is perhaps one of the most neat results of mathematical statistics: it is to reject the first 100/e % of the candidates and then follow the design given above. If one does this, the probability of getting the best candidate is also about 1/e for large n. Among the large literature on the Secretary problem, this author particularly recommends Freeman (1983) and Ferguson (1989).

Example 3. Allocation of treatments in clinical trials. Consider estimating the average rate of response in a clinical example with the model

$$E(X_i) = b_i\theta,$$

$$\text{Var } (X_i) = \ a \ \ (E(X_i))^p, \ \ p \geq 0$$

where X_i are independent observations on the response to b_i units of a stimulus. The goal is to estimate the response rate θ.

Suppose, as is often the case, a total fixed amount of the stimulus is available, and the dose that can be given to an individual has to be between two bounds (a very low dose is useless, and a very high dose is dangerous). Suppose the experimenter is unwilling to assume normality or any other model assumptions, and decides to estimate θ by using its BLUE.

The following is a design problem:

How many individuals should be used and how many doses should they get in order to minimize the mean squared error of the estimate?

It turns out the optimal design crucially depends on the value of p; in some cases, it is harmful to use more individuals for the study: i.e., a smaller sample is much better. In some other cases, the design is completely unimportant. and the same mean squared error is achieved regardless of the design. In some other cases still, it is best to use as many individuals as possible by applying the smallest amount of the stimulus. For an enjoyable look at this, one can see Kiefer (1987). A more recent generalization is DasGupta and Zen (1996).

Example 4. Sequential analysis. In sequential experiments, data come in a sequence, with the experimenter retaining the option of stopping and making an inference at any stage without collecting more data. The practical impetus for sequential experiments came from the frequency of real problems in which the inference problem was satisfactorily solved without the need of more data. In law enforcement, an analog may be that police stop taking tips incriminating other individuals when the existing evidence against a current suspect is overwhelming. In his monograph, Chernoff (1972) eloquently describes the relation of sequential experimentation to design: "in the act of deciding whether or not to gather more data, the statistician is making a choice of design. In this sense, sequential analysis ... is not a separate field (from optimal design)".

Deep and profound questions exist on what exactly is the "correct" optimal design problem in this context; particularly, the importance (or the lack of it) of exactly how the sampling process was terminated is a bitter bone of contention among statisticians of various descriptions. For a lucid and remarkably enjoyable discussion on this issue, one can see Berger (1986). Generally speaking, a strict believer in the Bayesian paradigm should have no use for knowing the exact termination rule; however, it is a truth that many declared Bayesians do not believe that with conviction. It is much like the parallel fact that randomization has no role in a strictly idealistic Bayesian world, but probably no Bayesians exist who would recommend against randomization, generally accepted to be a most sacred principle of statistical data gathering. Underneath all of these, the design problem is the following:

> Choose a stopping rule and a procedure for deciding between actions after one has stopped.

The exact optimal design in a strictly sequential context is usually not something one can write easily on a piece of paper; there are many instructive examples of Bayes optimal designs in sequential contexts in Chernoff (1972). Generally, one has to propose a design and establish its near optimality. Further contributions came from Schwartz (1962) and Siegmund (1985).

Example 5. Greedy algorithms. The Greedy algorithms refer to a whole family of optimization procedures in the problem of assigning k people to k jobs, when the cost of

assigning the ith person to the jth job is $C(i,j)$. Two common greedy algorithms are the following:

Method A. Assign to each individual the available job s/he does the best;

Method B. Initially, identify the best (i,j) combination; then eliminate the selected person and the selected job, and identify the next best (i,j) combination, and continue.

The design problem is the following:

Choose among the possible finite number of job assignments, the one that minimizes the total cost.

A mystifying result is that if $C(i,j)$ are *iid* Exponential, then each of Method A and B are equivalent to each other. The result is not trivial, and to see the equivalence one has to use various algebraic facts and other facts particular to the Exponential distribution, specifically, its memory-less property.

Example 6. Blackwell prediction. The Blackwell prediction algorithm deals with the problem of predicting the $(n+1)th$ member of an infinite $0-1$ sequence knowing the past members. A design thus corresponds to construction of an algorithm. This is probably one of the earliest examples of optimal design in which minimaxity appeared as a selection criterion (almost concurrently with the appearance of the Blackwell prediction algorithm, came the minimax ideas in Kiefer (1953), in which Kiefer uses minimaxity as a criterion for choosing a set of evaluation points in order to locate the maximum of a unimodal function in a bounded interval. Brown (1991) also gives a charming description of these growing years of optimal design).

Interesting things happen; a naive prediction algorithm (also appealing due to its simplistic nature) is to predict the $(n+1)th$ member as 1 if the average of the past members is $> 1/2$. While it predicts sequences that are really Bernoulli quite well, it does not do well for deterministic sequences of certain kinds. The Blackwell algorithm, which is randomized, in contrast seems to cover both types and has a minimax property. A Bayesian optimum design problem arises by putting a prior distribution on the unknown infinite sequence; but then, the problem is easily solved. One simply calculates the posterior probability of each value at the $(n+1)th$ stage and predicts the one with a larger posterior

probability. Blackwell gives a nice colloquial description of the relevance of this problem in Information theory and artificial intelligence in his interview with Morris DeGroot in Statistical Science (1986).

Example 7. Computer experiments and infinite dimensional problems. In recent years, emphasis in optimal design is shifting to new families of problems. One such area involves the writing of a stochastic equation for predicting the output of a deterministic computer code. Typically, the problem is of the following type:

One has a (possibly very high dimensional) input x in response to which the computer produces an output $y = y(x)$. The relation is supposed to be deterministic. However, a stochastic model is introduced:

$$y(x) = \text{ A regression function } + z(x),$$

where $z(x)$ is an error. The idea is to build a predictor by treating this as a regression problem; this predictor acts as a (cheaper) proxy to the deterministic computer output of the complex code. The design problem is the following:

> Determine a set of n values of x which are to be used in constructing the prediction equation.

As in Example 3, one can decide on a linear method and avoid making assumptions about the stochastic process $z(x)$ or one can assume $z(x)$ is a path of a certain well understood process, typically a Gaussian process. Linear estimation in this context is commonly called kriging; one can use a kriging procedure together with a specified design criterion to arrive at a functional that needs to be maximized to construct an optimal design. Common design criteria include an integrated mean squared error (over x, with respect to some probability measure on x), and a maximum mean squared error. The problems are much harder than what one sees in ordinary regression designs; as a consequence, it is typical that the optimal design has to be found by a search method and the construction of the search algorithm is as important as the identification of the criterion functional. One can see Sacks, Welch, Mitchell and Wynn (1989) and Sacks and Schiller (1988) for a broad exposition. Use of stochastic processes as priors on continuous functions is also done in Diaconis (1987), O'Hagan (1978), among others. Essentially the same things also go by the name of illposed inverse problems.

The above examples clearly show how rich the study of optimal designs is and can be. It is an error to think that optimal design is an abstract area of mathematical statistics limited to standard statistical models like factorial experiments, or linear and nonlinear models. In fact, according to this author, the more lively optimal design problems arise in branches outside of these models, and much remains to be looked at. As commented earlier, a remarkably vast literature already exists on optimal design. Anyone seriously interested in learning about optimal design must at the least consult , in addition to the references above, Kiefer (1959), Atkinson and Donev (1992), Fedorov (1972), Silvey (1980), Pukelsheim (1993), Pilz (1991), and in particular for Bayesian optimal design, a very recent review article by Chaloner and Verdinelli (1994). Indeed, we will make a conscious effort to as much as possible emphasize aspects and literature not emphasized in this review article in order to avoid a wasteful duplication of intellectual effort. However, there will necessarily be some overlapping due to the review nature of both articles.

2. Outline.

In Section 3, we give some history of the theory of optimal design; naturally, this will include the early developments usually attributed to Kiefer and Wolfowitz (1959). In Section 4, we explicitly start to discuss Bayesian formulations of the design problem, and we will discuss the direct impact of the Kiefer-Wolfowitz theory on Bayes design and also discuss the early history of Bayes design. In Section 5, we will broadly discuss the typical mathematics of Bayes regression designs; this will cover the Elfving theorem, other geometry due to Chaloner (1984), and El-Krunz and Studden (1991), and Studden and Dette (1993), and will also include the role of moment methods and equivalence theorems. Section 6 will apply the theory to explicit description of Bayes optimal designs; this section will also include some work on Factorial experiments, in particular those of Notz,Toman, and their coauthors. In Section 7, we critically assess the relevance and impact of optimal design theory, and address issues such as belief in the model, and construction of all around designs. This section will include an outlook into how optimal design theory can adapt itself to the opinion of practitioners. Section 8 discusses the nonconjugate case, and whether Bayesians need to even worry about optimal designs : this will cover two aspects - whether classical designs alone suffice, and whether the prior matters. Robust Bayes optimal designs will be discussed in this section. Section 9 will cover miscellaneous design

problems, as in quality control, engineering reliability, spatial designs, etc. Section 10 covers sample size and preposterior formulations of the Bayes design problem as opposed to standard criteria like integrated Bayes risk. Section 11 gives a brief exposition to nonlinear models and associated design problems. In Section 12, various other issues are discussed and concluding remarks are made. Section 13 contains an appendix on the mathematical tools of optimal design.

3. Some history.

Optimally setting up an observational study so as to extract as much relevant information as possible is such a natural idea that it is possibly impossible to trace back to the first scientific article on this; there are, however, some demonstrably early ones. Smith (1918) already has a clear flavor of optimal design in polynomial regression; apart from one intermediate but clearly a key contribution by Wald (1943), the culture of a structured optimal design theory arrived with Jack Kiefer. The paper by Wald (1943) was key in its influence on how optimal design theory was formulated and was done for three decades. Although the time around the second world war was in some sense the golden age of decision theory (every eminent statistician did some decision theory around that time), it is fair to think that Wald's 1943 article had a binding influence on the formulations that came through later. Stigler (1974) gives a fine account of the history of polynomial regression that makes interesting reading for researchers in design of experiments.

The most remarkable and time tested contribution of the Kiefer - Wolfowitz theory was the concept of an approximate design. Indeed, this concept had such a tremendous impact that optimal regression design is done even today more or less within the domain of approximate designs. The idea was that exact optimal designs for a given sample size n are to a significant extent dependent on the value of n, and their derivation corresponds to a straightforward (but not proportionately enlightening) integer programming problem; in contrast, by formulating a design as a probability measure on the design space, two things are achieved: avoiding a dependence on n (except at the implementation stage) and making possible a strikingly beautiful theory that connects together several branches of mathematics (analysis in particular); Karlin and Studden (1966) is a standard reference on connections of optimal design theory to moment methods and orthogonal polynomials. The other important contributions of the Kiefer-Wolfowitz theory were the concepts of

alphabetic optimality; these are accepted quite universally as criteria for evaluation even today, although a somewhat small school has argued against a few criteria in use due to their apparent lack of correspondence to decision theory based on a utility function: see Chaloner and Verdinelli (1994) for more on this.

An important point at which regression and other (factorial or qualitative) types of optimal design theory separated is the adoption of approximate designs as a founding concept. Historically, in these other branches of optimal design theory, combinatorics and integer programming continued to play the key roles. Bose (1948) is an early important work followed by much work of many researchers, notably C.S. Cheng. One can see Shah and Sinha (1989) for a comprehensive and informative account of nonregression optimality theory; Kurotschka (1978) gives an account of the nature of optimality theory in the presence of both quantitative and qualitative factors.

A mathematical tool from which all of optimal design theory benefitted is commonly called an equivalence theorem. At a basic level, an equivalence theorem only states that at a point of minima a differentiable function has derivative zero and it increases as one moves away from the minima in a given direction. The use of an equivalence theorem is in its ability to verify that a design suspected to be optimal is indeed so; subject to numerical accuracy of such a verification, this has been creatively used in a number of problems, notably in nonlinear models by Kathryn Chaloner and her coauthors. Statements of general equivalence theorems can be seen in many writings; one can see in particular Silvey (1980), Pukelsheim (1993), and Schoenberg (1959).

As much as the Kiefer-Wolfowitz theory was beautiful, its impact on practitioners was limited. The problem is in the nature of the optimal designs. One has to trust the model absolutely to consider actually using these exact designs. It is therefore quite natural that concerns about robustness with respect to misspecification of the model were voiced; Stigler (1971) and Studden (1982) reacted to these concerns, among many others. There are also many who believe that parameter estimation is not the aim of an experiment and a model should be assessed on the basis of its predictive power; it is a fairly persuasive argument and not surprisingly, predictive design criteria have been suggested. Lindley (1968) is probably the earliest article on Bayesian-decision theoretic design based on predictive evaluations;

the topic continued with a number of articles by Brooks (1974, 1976), and has recently gained further momentum with an article due to Eaton, Giovagnoli and Sebastiani (1994).

The history of a structured Bayesian optimality theory is by far much more recent; in fact, it can be said that Chaloner (1984) is the first serious attempt to develop a theory of Bayes optimality in linear regression context. Chaloner (1984) gives a formulation, shows analogs of the Elfving geometry of the classical theory in some special cases and explicitly describes the role of the prior on the difference between the Bayes and classical optimal designs. Meanwhile, a number of people in Europe started to actively work on Bayesian optimal designs, and Pilz (1991) is an early contribution that assembled a great amount of material in the context of linear regression and really provided a solid impetus for further work. The Elfving geometry in the context of Bayes designs was beautifully described in El-Krunz and Studden (1991), perhaps the deepest theoretical contribution to Bayes designs till now. Sensitivity of Bayes designs to the choice of the prior was given a structured formulation and explicit robust Bayes designs were given in DasGupta and Studden (1991); Toman (1992) treats sensitivity in models with qualitative factors.

It was already well known that practically all of the neat theory of optimal designs one sees in linear models is unachievable in even the simplest kinds of nonlinear models. In fact, a great amount of philosophical and moral dilemma pervade optimal design in nonlinear models. The problem is that strictly speaking, an optimum design depends on the true value of the parameter one is trying to estimate in the first place. The concept of local optimality was introduced in the classical theory to tackle this issue. The first attempt at seriously working out Bayesian optimal designs in a series of nonlinear models was made in Chaloner and Larntz (1989), although prior important contributions exist, notable among them Box and Lucas (1959). The mathematics of the Bayesian optimality theory for nonlinear models is challenging, and rather surprising advances have come through in a short period of time. The works of Holger Dette and his coauthors deserve specific mention due to their insightful nature. Unfortunately, however, all the evidence still suggests that a unifying theory as in the case of linear models would not be possible and a piecemeal theory may emerge with time.

The theory and practice of Bayesian optimal design are still at an early stage; many

topics for which a great amount of results exist under classical optimality criteria have not been at all looked at. Effect of dependence in the observations is one such (old works of Sacks and Ylvisaker (1966) and Bickel and Herzberg (1979) are by now classic contributions to this) topic. Determination of minimum and optimal sample size which has taken the status of textbook material in classical statistics, is just beginning to get serious attention in terms of a debate about which formulations are proper for the Bayesian; some early theoretical works include DasGupta and Mukhopadhyay (1994) and DasGupta and Vidakovic (1994). It is encouraging to see that efforts are being made , although somewhat in isolation of a structured theory, to work out Bayesian optimal designs in actual applied problems; workers at the Duke school have already made good contributions in this area. Caselton and Zidek (1984) and Schumaker and Zidek (1993) are instances of elegant theoretical developments in interesting real problems. Another area in which some effort is being made is the writing of computer codes for numerical implementation of Bayes optimal designs (Clyde (1993)). Significant literature on this already exists for construction of classical optimal designs; in particular, one can see Atkinson and Donev (1992) for an exchange algorithm much like the exchange algorithm of numerical analysis for finding the minimax fit to continuous functions from Haar spaces, and Haines (1987) for an innovative use of simulated annealing in constructing D-optimal designs.

For comprehensive reading of optimal design, both classical and Bayes, many excellent sources exist; we enthusiastically recommend Atkinson and Donev (1992), Box and Draper (1987), Herzberg and Cox (1969), Pukelsheim (1993), Silvey (1980), Chaloner and Verdinelli (1994), and Wynn (1984) for anyone interested in this topic. In fact, our effort would be to emphasize whenever possible specific points not addressed in much detail in these earlier contributions. It is also necessary to consult these for bibliography in addition to the bibliography of this article.

4. Alphabetic criteria and other formulations.

4.1. Approximate designs. The five most widely accepted criteria for an optimality theory of designs are c, A, D, E, and G optimality; there are others. The road to arrival at these criteria can be thought of in the following way: one has a standard Gauss-Markov linear model and decides to use the leastsquares estimate of the regression coefficients; it seems natural that one should want to make the estimate as accurate as possible. Since

the leastsquares estimate is already unbiased, consideration will then focus on the variance covariance matrix of the leastsquares estimate. Minimizing the trace, determinant, and the maximum eigenvalue of this matrix respectively correspond to A, D, and E optimality. Minimization of the variance of the leastsquares estimate of a linear combination of the regression coefficients corresponds to c optimality. G optimality, which corresponds to an average over linear combinations of the coefficients thus also links up to essentially the same fundamental idea.

Consider then the usual linear model $y_i = \sum_{j=0}^{p} \theta_j \ f_j(x_i) + \epsilon_i$, where the vector of errors satisfies $E(\underset{\sim}{\epsilon}) = 0$ and $D(\underset{\sim}{\epsilon}) = \sigma^2 I$, where $\sigma^2 > 0$ is possibly unknown. A large number of statistical models in everyday use fall under this general setup, and in principle, therefore, an optimality theory for designs in the canonical linear model certainly has a wide scope for application. The leastsquares estimate for $\underset{\sim}{\theta}$ has the representation $(X'X)^{-1}X'Y$, with variance covariance matrix $\sigma^2 \ (X'X)^{-1}$, where X denotes the design matrix with rows $(1, f_1(x_i), \ldots, f_p(x_i))$. These statements need to be slightly changed when $X'X$ is not full rank, which in fact does happen in some interesting problems. We shall later see that usually this ceases to be a problem in the corresponding Bayes theory. For the alphabetic criteria listed above, one can make a transition to the precision matrix $\frac{1}{\sigma^2}(X'X)$, due to the well known relations between the trace, determinant and eigenvalues of a matrix and its inverse. Actually, there is a whole family of criterion functions that permit such a transition from the "dispersion" matrix $(X'X)^{-1}$ to the "information" matrix $X'X$. Thus, for instance, the E optimality criterion corresponds to maximizing the minimum eigenvalue of the information matrix $X'X$.

Now if the distinct rows in the "design" matrix X are denoted as $\underset{\sim}{x}_1', \underset{\sim}{x}_2', \ldots$, with multiplicities $n_1, n_2, \ldots$ (a repeated row corresponds to replication of the same levels of the independent variables for two or more individuals), then the information matrix takes the form

$$X'X = n \sum n_i/n \ \underset{\sim}{x}_i \ \underset{\sim}{x}_i';$$

Writing n_i/n as p_i, one therefore sees that $X'X$ equals an average of the quantities $\underset{\sim}{x}_i \underset{\sim}{x}_i'$. The idea of an approximate design is to allow an arbitrary probability measure instead of a discrete probability vector $\underset{\sim}{p}$ such that the elements of $n.\underset{\sim}{p}$ are integers. One then has a

general information matrix

$$M = M\ (\mathcal{E}) = \int \underset{\sim}{x}\underset{\sim}{x}'\ d\ \mathcal{E}(x),$$

where $\mathcal{E}$ is a probability measure on the design space $\mathcal{X}$.

Example 1. Consider quadratic regression $y = \theta_0 + \theta_1 x + \theta_2 x^2$, with x varying in the interval $= [-1, 1]$. Then the information matrix M is immediately seen to be

$$M = \begin{bmatrix} 1 & c_1 & c_2 \\ c_1 & c_2 & c_3 \\ c_2 & c_3 & c_4 \end{bmatrix},$$

where c_i denotes the ith moment of the probability measure $\mathcal{E}$. Notice the interesting fact that the elements of M are moments of $\mathcal{E}$. Because of this reason, it is quite common to refer to the information matrix as a moment matrix as well.

This connection of information matrices to moments also helps illustrate the form of optimal designs according to the alphabetic criteria described above in polynomial regression. For instance, according to the D-optimality criterion, one should try to identify a probability measure on $[-1, 1]$ such that the corresponding moment matrix has a maximum determinant. The original (finite dimensional) integer optimization problem has now changed to an infinite dimensional problem on the space of probability measures. Ironically, this complexity actually adds structure and simplicity to the problem. Assuming for a moment, that a probability measure giving a largest value of the determinant exists, it is clear that this particular $\mathcal{E}$ must give the largest value of the moment c_4 among all probability measures which produce the same values as those of $\mathcal{E}$ for the lower moments c_i, $i = 1, 2, 3$. Theorems in moment theory and an easy symmetry argument now imply that there is an optimal choice of the probability measure with supports at $0, \pm 1$ and calculus then shows that the weight at 0 has to be $1/3$. The solution to the D-optimal problem for quadratic regression on $[-1, 1]$ according to the approximate design theory is thus to take an equal number of observations at $0, \pm 1$. Of course, if the total sample size is not a multiple of 3, the ideal design has to be rounded to an integer design; even more, even if the total sample size was a multiple of 3, the ideal D-optimal design from the approximate theory need not coincide with the solution that would obtain from the integer problem. These are issues one needs to be aware of, but the strong structure that

the approximate theory provides more than makes up for these somewhat minor issues. A much more serious issue is that the D-optimal design is too thinly supported; in view of this, it is rare for an exact optimal design to be religiously used in practice. But they provide a useful yardstick for the performance of other designs under the assumption of an approximate validity of the regression model. Pukelsheim (1993) and Pukelsheim and Rieder (1992) write eloquently about these issues.

All common criteria ϕ for optimal design satisfy a monotonicity property in the information (moment) matrix M : if one considers two such matrices M_1, M_2, with $M_2 \geq M_1$ in the Loewner ordering, i.e., if $M_2 - M_1$ is nonnegative definite, then the criterion ϕ satisfies $\phi(M_2) \geq \phi(M_1)$. This motivates the following definition:

Definition. An information matrix M_1 is called inadmissible if there exists another information matrix M_2 such that $M_2 > M_1$ (i.e., $M_2 - M_1$ is nonnegative definite but not the null matrix).

In construction of optimal designs, it is therefore necessary to only consider probability measures resulting in admissible information matrices: this is like the well known fact in decision theory that admissible rules form a complete class. In polynomial regression problems, due to the moment interpretation of the information matrix, this helps in bounding the number of support points in an optimal design according to any criterion that is monotone increasing in the moment matrix in the Loewner ordering. Indeed, the following holds:

Theorem. Under the hypothesis of monotonicity of ϕ in the Loewner ordering, an optimal design for a polynomial regression model of degree p can have at most $p+1$ points in its support with at most $p-1$ points in the interior of $\mathcal{X}$.

This result aids in understanding why the theoretical optimal designs are generally so thinly supported. Further pinpointing of the exact number of points and their weights do not come out of this theorem.

4.2. Bayesian formulation of an optimal design problem. In a strictly Bayesian decision theoretic setup, one has a set of parameters θ with a prior distribution G, a specified likelihood function $f(x|\theta)$, and a loss function $L(\theta, a)$. Given a design, there

is an associated Bayes rule with respect to the trio (f, L, G); an optimal design should minimize over all designs the Bayes risk , i.e., the average loss of the Bayes estimate over all samples and the parameters. Chaloner and Verdinelli (1994) give a fairly comprehensive review of this formulation. In particular, they give a number of loss functions that have been proposed, and there is an instructive account of which alphabetic Bayesian criteria correspond to such a a loss-prior formulation. Note that the formulation can as well take the route of prediction rather than estimation; Eaton, Giovagnoli and Sebastiani (1994) consider a predictive formulation and show that sometimes one returns with the alphabetic optimal designs again, but not always.

There is another (simplistic) way to look at the Bayes design problem which in fact has the axiomatic justification under a normal-normal-gamma linear model with squared error loss. Thus, consider the canonical linear model $\underset{\sim}{Y} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, $\underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2 R^{-1})$. Then, under the standard squared error loss $\|\underset{\sim}{\theta} - \underset{\sim}{a}\|^2$, the Bayes risk (in fact even the posterior expected loss itself) equals $tr(M + R/n)^{-1}$, where n denotes the sample size. One would therefore seek to minimize $tr(M + R/n)^{-1}$, which has a remarkable similarity to the classical A-optimality criterion.

The Bayesian alphabetic criteria are thus defined for linear models as:

Bayesian A-optimality: Minimize $tr(M + R/n)^{-1}$,

Bayesian D-optimality: Minimize $|M + R/n|^{-1}$,

Bayesian c-optimality: Minimize $c'(M + R/n)^{-1}c$ for a given vector c,

Bayesian E-optimality: Minimize the maximum eigenvalue of $(M + R/n)^{-1}$,

Bayesian G-optimality: Minimize $\int c'(M + R/n)^{-1}c \, d\nu(c)$, where ν is a probability measure on the surface of the unit ball $c'c = 1$. (note that Studden (1977) calls this integrated variance optimality).

Of course, in the absence of a meaning for R, these criteria do not stand to reason. They do stand to reason by doing one of two things: a structured setup of normal-normal-gamma distributions with a squared error loss, or restriction to affine estimates with only assuming that the dispersion matrix of $\underset{\sim}{\theta}$ equals $\underset{\sim}{\sigma}^2 R^{-1}$. The presence of σ^2 as a factor in

the dispersion matrix of $\underset{\sim}{\theta}$ makes this less inoccuous than it seems.

A substantial amount of the optimality theory in Bayes design has been done with these alphabetic criteria. Note that if the sample size n is even reasonably large, the extra factor R/n in these functionals should not (and indeed do not) play much of a role. Thus, for priors in linear models which are not flatter in comparison to the normal likelihood tend to report optimal Bayes designs that track the classical ones very closely, or even exactly. On the other hand, although there is some scope for optimality work with t or other flat priors, so far there are no published works in this direction. The field of Bayes optimal designs therefore still holds out some (hard) open problems even for the Gauss-Markov linear model.

Of course, estimation and prediction are not the only inference problems one can design for; indeed, the design to be used should be consistent with what would be done with the data. The role of optimal designs in testing problems is described in Kiefer (1959), where he shows that for maximizing the minimum power over small spheres around the null value in ANOVA problems, it is not correct to use the F test regardless of the design. Kiefer's criterion would not be very interesting in a Bayesian framework (although some Bayes design work has used average power as the criterion: see Spiegelhalter and Freedman(1986)); however, Bayes optimal design for testing problems has generally remained neglected. DasGupta and Studden (1991) give a fully Bayesian formulation and derive Bayes designs; there are also a number of remarkably charming examples in Chapter 7 of Berger (1986), and there is some more theory with conjugate priors in normal linear models in DasGupta and Mukhopadhyay (1994).

In closing, the Kiefer-Wolfowitz theory has had a profound impact on the work in Bayes optimal designs in two ways: use of the alphabetic criteria and adoption of the approximate theory.

5. Mathematics of Bayes design.

5.1. General exposition. The mathematics of Bayes optimal designs is generally the same as that in classical optimal design. There are three main routes to obtaining an optimal design: *i.* use an equivalence theorem, *ii.* In polynomial models, use inherent symmetry in the problem (if there is such symmetry) and convexity of the criterion func-

tional in conjunction with Caratheodory type bounds on the cardinality of the support, and *iii*. Use geometric arguments, which usually go by the name of Elfving geometry, due to the pioneering paper Elfving (1952).

An equivalence theorem does the following: it prescribes a function $F(\mathcal{E}, x)$ defined on the design space such that $F(\mathcal{E}, x) \leq 0$ for all x in $\mathcal{X}$ and is $= 0$ if and only if x is in the support of an optimal design $\mathcal{E}$. Usually, but not always, some guess work and some luck is involved in correctly using equivalence theorems for identifying an optimal design. The nice thing about equivalence theorems is that really general equivalence theorems are known that cover probably almost all cases one would be interested in, and in principle, it is supposed to work. One can see Silvey (1980), Whittle (1973) and Pukelsheim (1993) for increasingly general equivalence theorems.

Convexity arguments do the following: First by using Caratheodory type theorems, or if possible upper principal representations from moment theory, one gets an upper bound on the number of points in the support of an admissible design. Then, one proves that the criterion functional has some symmetry or invariance property; finally, one proves that the functional is convex in a convex class of moment matrices. Application of all of these together would reduce the dimensionality of the problem to a very low dimension, which is then solved by standard calculus.

The geometric methods attributed to Elfving (and developed by many others subsequently) are by far the most subtle methods of optimal design theory, and need to be stated very carefully with changes in the criterion function. It is best understood by a verbal geometric description for the c-optimality problem. For this, one takes the symmetric convex hull of the design space, i.e., $E = CH(\mathcal{X}U - \mathcal{X})$, where CH denotes convex hull. This set is symmetric, convex and compact provided $\mathcal{X}$ is compact. Now take any vector $\underset{\sim}{c}$; if $\underset{\sim}{c} \neq \underset{\sim}{0}$, then on sufficient stretching or shrinking, it will fall exactly on the boundary of the convex set E (the scalar by which $\underset{\sim}{c}$ is divided in order that this happens is called the Minkowski functional of E evaluated at $\underset{\sim}{c}$). Call this scaled vector $\underset{\sim}{c}^*$. Then $\underset{\sim}{c}^*$ can be represented in the form $\sum p_i y_i$ where each y_i is either in $\mathcal{X}$ or $-\mathcal{X}$. If $\mathcal{X}$ is not already symmetric, then those that are in $\mathcal{X}$ give the support of an optimal design. A concise general version of this method for c-optimality is given in Pukelsheim (1994); there is also

a wealth of information with many greatly unifying results in Dette (1993). One should be cautious about the use of the terminology "prior" in Dette (1993); the unifying nature of the theorems is the most gratifying aspect of this article, but the worked out examples indicate that again elements of intuition and good luck are needed for the Elfving geometry to be useful.

5.2. State of the art in Bayesian alphabetic optimality.

5.2.1. c-optimality. It seems that the best results on Bayesian optimality are known for this criterion. Chaloner (1984) already considers Bayesian c-optimality and gave a form of the Elfving geometry in this case. Her results imply that Bayesian c-optimal designs can be one point, i.e., they can sometimes take all observations at one point. The deepest results on Bayesian c-optimality are given in El-Krunz and Studden (1991). They succeeded in achieving the following:

(a) give a characterizing equation completely specifying a c-optimal design, together with a Bayesian embedding of the classical Elfving set that describes the c-optimal design,

(b) characterize the situations when the c-optimal design is in fact one point,

(c) characterize the situations when a particular one point design is c-optimal,

(d) characterize the cases when the classical and the Bayesian c-optimal designs are exactly the same, and

(e) demonstrate that for any prior precision matrix, there is a sufficiently large sample size beyond which the classical and the Bayesian c-optimal designs have exactly the same support. This last result has a remarkable consequence: it is a classic fact (see Karlin and Studden (1966)) that in polynomial regression, for the extrapolation problem, i.e., for estimating the mean response at an x outside of the design space, the c-optimal design is always supported at the same set of points (it is a particularly brilliant application of the methods of orthogonal polynomials to optimal designs). Therefore, the result in El-Krunz and Studden (1991) demonstrate the same property for the Bayesian c-optimal design in the extrapolation problem for any prior precision matrix provided the sample size is large.

This is extraordinary, because one is saying much more than weak convergence to the classical design. That the supports coincide for large sample sizes was already recognized in Chaloner (1984) also.

5.2.2. A-optimality. The criterion for c-optimality can be written in the equivalent form $tr(cc'(M + R/n)^{-1})$. A generalization of this is the functional $tr(\varphi(M + R/n)^{-1})$, where φ is some nonnegative definite matrix of rank $k, k \leq p$, where p denotes the rank of the information matrix M. This corresponds to Bayesian A-optimality.

A general equivalence theorem for this case is given in Chaloner (1984); a clean version of this equivalence theorem is available in Dette and Studden (1994b). In principle, this theorem can be used to find a φ-optimal Bayes design, but as with all equivalence theorems, one almost has to guess the design to use the theorem. This is like theorems in minimax theory which provide excellent vehicles for verifying that a good guess is in fact a minimax rule, but are not tremendously useful in guiding to the rule. In fact, a subsequent result in Dette and Studden (1994b) is of much greater practical use: in this result, they show that phenomena earlier described for Bayesian c-optimality continue to hold for φ-optimality. This result completely describes a value of a threshold sample size after which the Bayes and classical optimal designs are identically supported and then even gives the weights at the support points for the Bayes design.

For the case when φ is full rank, i.e., $k = p$, the invariance-convexity arguments outlined in section 5.2.1 can be used for certain types of prior precision matrices. One can see DasGupta and Studden (1991) for some further hints on this. In general, however, calculus followed by application of Caratheodory type bounds appears to be the only method that will apply. For bounds on the number of points in the support of the Bayesian A-optimal design that are improvements on the Caratheodory bounds, one should see Theorem 2 in Chaloner (1984).

5.2.3. D-optimality. Dykstra (1971) has a flavor of Bayesian D-optimality; the theoretical foundation seems to be the convexity arguments presented in DasGupta and Studden (1991). It would be nice to find out if the classical and Bayes D-optimal designs share the same kinds of properties as they do for c-optimality. Although in regression designs either theory or explicit examples seem to be lacking, there is some work on Bayesian

D-optimality for factorial designs and also for nonlinear models. We will discuss these in subsequent sections.

5.2.4. E-optimality. The state of the art results in Bayesian E-optimality for the canonical linear model are again in Dette and Studden (1994b). There is an equivalence theorem; however, we do not recommend trying to use it. The useful results are remarkable in their neatness. There are two such results: one asserts that for sufficiently large samples, the Bayes and the classical E-optimal designs are identically supported and gives a formula for calculating the weights. This calculation can be daunting. A second result asserts that under two conditions the classical E-optimal design is exactly identical to the Bayes solution. In general, these hypotheses are not easy to verify. However, for the important case of polynomial regression, they show a clean relation to an earlier result of Pukelsheim and Studden (1993) for classical E-optimal designs. Pukelsheim and Studden (1993) showed that the classical E-optimal design is supported at the points of peak of the pth Chebyshev polynomial $T_p(x)$, which are $\{-cos(j\pi/p), 0 \leq j \leq p\}$ for the interval $[-1, 1]$, where p as before denotes the degree of the polynomial regression model. Dette and Studden (1994b) show that the Bayes E-optimal design is supported at these same points and in addition give an easily computable equation for the weights. This result is valid for sufficiently large samples.

5.2.5. Implications in practice. The results we see in sections 5.2.1-5.2.4 demonstrate two things: first, for practically every one of these alphabetic criteria, exact identification of a Bayes optimal design is at least a time consuming process, despite the fairly good theory that already exists. Second, if one is willing to use a conjugate prior in a Bayes formulation of the design problem, then use of at least the same support points as the corresponding classical design is wise. One can and probably should do a numerical search to find out if the same weights can be used without much harm as well; the same search should help locating better weights if indeed there are much better weights than the classical ones. In this sense, the results described above are tremendously valuable. They show an overwhelming structure, and demonstrate that subject to using these alphabetic criteria, and conjugate priors, trying to exactly identify a Bayes design is not a particularly good idea. Of course, in the case of very small samples, prior information is more important, and the results stated earlier are not valid!

6. Examples and other information of use to practitioners.

6.1. General examples. To get a flavor for how the theorems of Bayes optimality theory apply under the various alphabetic criteria, one can benefit from examples described in Brooks (1976), Gladitz and Pilz (1982), Chaloner (1984), DasGupta and Studden (1991), El-Krunz and Studden (1991) and Dette and Studden (1994b). These are for what are commonly called regression designs. Analogous examples for other kinds of models would be cited in later sections.

6.1.1. Regression in a sphere: Suppose the design space is the sphere $\{\underset{\sim}{x} : \sum x_i^2 \leq 1\}$. This can thus be regarded as an example of multiple linear regression without an intercept term. Chaloner (1984) gives several examples of φ-optimal Bayes designs in this case. El-Krunz and Studden (1991) also consider regression in a sphere and have an example, which has a flavor of a theorem. Brooks (1976) adjusts the spherical design space in order to entertain an intercept term. Both Chaloner (1984) and Gladitz and Pilz (1982) address the issue of rounding the optimal design to an implementable integer design and do real numerical examples on the associated loss of efficiency.

6.1.2. Regression in a cube or on a discrete set of points. Multiple linear regression in which each independent variable lies within $\pm a$ for some $a > 0$ corresponds to regression in a cube. Regression on a discrete set of points is an important example for many physical, chemical and environmental experiments.

Chaloner (1984) gives an example of the application of an equivalence theorem involving optimality for regression in a cube; El-Krunz and Studden (1991) give an illuminating example of regression on a set of three points. This example brings out all the important features of the Bayes optimality theory: that for large samples, the Bayes design coincides with the classical, and sometimes they are at least identically supported and for very small samples, prior information is more important and the Bayes design is not even supported on the same points as the classical solution. We recommend this example to everyone.

6.1.3. Polynomial regression. Certainly this is the case in which the maximum number of worked out examples are available. DasGupta and Studden (1991) give an example of Bayesian D-optimal designs for quadratic regression. Chaloner (1984) and El-Krunz and Studden (1991) both give the example of estimating the coefficient of x^3 in cubic regression.

There is also an example involving quadratic regression in El-Krunz and Studden (1991) that illustrates the subtle geometry of the Elfving method for the Bayesian theory. Dette and Studden (1994b) have a number of examples on E-optimality for polynomial regression. In one of these examples, they apply their theorems to show a remarkable phenomenon: although the Bayes designs are supposed to depend on the sample size for small n, in certain instances a very small sample size may suffice for the Bayes optimal design to already coincide with the classical design. There is another example in particular where they obtain for quadratic regression the Bayes E-optimal design as n changes, with an arbitrary prior dispersion matrix.

The use of these concrete examples is in two aspects: someone interested in alphabetic Bayes designs can get a feeling for the theory by seeing it applied, and also get a feeling for which aspects are important, namely the prior dispersion matrix or the sample size, etc.

6.2. Exact classical designs in some important cases. Since Bayes and classical optimal designs tend to be either exactly the same or very similar under alphabetic criteria whenever one uses conjugate priors and the sample size is not very small, for practicioners (Bayesian or not) it is greatly useful to know the classical optimal designs in some important cases. Again, it is the view of most researchers in this area that optimal designs are not intended for religious use, but are to be used as standards of evaulation for other designs. In the following, we give the classical optimal designs for polynomial regression when the single independent variable belongs to the symmetric interval $[-a, a]$. We can take $a = 1$ and scale the design if a is different from 1.

6.2.1. A-optimality. This corresponds to minimizing the trace of the dispersion matrix of the least squares estimate. The optimal designs are as follows; in each case the symmetric members of a pair have equal weight and the weights are in the same sequence as the

points. The table is taken from Pukelsheim (1993), where more information is available.

Degree of Polynomial	Points in Support	Weights
1	± 1	.5
2	$0, \pm 1$	.25, .5
3	$\pm .464, \pm 1$	.349, .151
4	$0, \pm .677, \pm 1$	.29, .25, .105
5	$\pm .291, \pm .789, \pm 1$	.232, .188, .08
6	$0, \pm .479, \pm .853, \pm 1$	.205, .185, .148, .065

6.2.2. D-optimality. The classical D-optimal designs for polynomial regression have a property that are regarded by some as bad and others as good: if there are k points in its support then each point has weight $1/k$. Thus a D-optimal classical design is uniform on its support. The optimal designs are as follows; the points in support are just the turning points of the pth Legendre polynomial plus the endpoints.

Degree of Polynomial	Points in Support	Weights
1	± 1	equal
2	$0, \pm 1$	1/3 each
3	$\pm .447, \pm 1$	1/4 each
4	$0, \pm .655, \pm 1$	1/5 each
5	$\pm .285, \pm .765, \pm 1$	1/6 each
6	$0, \pm .469, \pm .830, \pm 1$	1/7 each

Again, an extended version of this table can be seen in Pukelsheim (1993).

6.2.3. E-optimality. Pukelsheim and Studden (1993) proved the following general result on classical E-optimal designs for polynomial regression on $[-1, 1]$:

The classical E-optimal design is supported on the points $\{cos(j\pi/p), 0 \leq j \leq p\}$, where p denotes the degree of the polynomial. Furthermore, the weights $\{p_i\}$ are proportional to $(-1)^{p-i}u_i$ where $\{u_i\}$ satisfy the system of linear equations $\sum u_i(cos(j\pi/p))^i = c_j$, where c_j is the coefficient of x^j in the pth Chebyshev polynomial $T_p(x)$. For instance, the fourth Chebyshev polynomial is $T_4(x) = 8x^4 - 8x^2 + 1$, and therefore the coefficients $\{c_j\}$ are respectively 1, 0, −8, 0, 8. In view of this general result, it is not difficult to write the classical E-optimal designs for polynomial regression. They are as follows:

Degree of Polynomial	Points in Support	Weights
1	±1	.5
2	0, ±1	.6, .2
3	±.5, ±1	.373, .127
4	0, ±.707107, ±1	.318, .248, .093
5	±.309017, ±.809017, ±1	.246, .180, .074
6	0, ±.5, ±.866025, ±1	.218, .189, .141, .061

6.2.4. c-optimality. The classical c-optimal design can be found, in principle, by using the Elfving method, for any given vector c. In particular, if $c = (1, x, x^2, \ldots, x^p)$ for $|x| > 1$, one has the problem of "extrapolation" whereas if $|x| \leq 1$, one has a problem of interpolation. Even these two cases are radically different in the corresponding optimal designs. In the first case, regardless of the value of x, the optimal design is supported at the points given in section 6.2.3, while in the latter case, the optimal design has only the given value x in its support. Optimal designs for the individual coefficients in the various powers of x can be written down. The solution depends on whether the subscript of the coefficient is an even or odd integer away from p, the degree of the polynomial model. Thus there is no universal statement one can write on a piece of paper for easy communication. One can see Pukelsheim and Studden (1993) or Pukelsheim (1993) for further information, if needed.

6.3. Factorial models, treatment control comparisons, elimination of heterogeneity and ANOVA.

6.3.1. Factorial models. The first attempt at derivation of Bayes optimal designs for factorial models seems to be Owen (1970).Work following this includes Smith and Verdinelli (1980), Verdinelli (1983), Giovagnoli and Verdinelli (1983), Toman (1987), Hedayat, Jacroux and Majumdar (1988) and Toman and Notz (1991). A recent important and comprehensive contribution is Majumder (1995). Generally, the work has derived alphabetic optimal Bayes designs for some selected models. These models include one and two factor models, one and two way ANOVA, and one and two way heterogeneity models. Almost exclusively, these works have assumed multivariate normal priors in conjunction with multivariate normal observations, which are needed for analytical results. Some of these works also consider the computational aspects, and give special cases where the computation simplifies. In some of these models, there have been some work on sensitivity with respect to the prior distribution which would be cited later.

6.3.2. A-optimality. In the two factor ANOVA model, Owen (1970) derives a general result giving the optimum allocation of treatments given a specified blocking. Owen's criterion should really be called generalized A-optimality due to the general quadratic loss he has for estimating the treatment effects. Owen shows that in certain cases, the computational aspect simplifies. A-optimality is also considered in Giovagnoli and Verdinelli (1985) in one way heterogeneity models and in Hedayat et al for two way heterogeneity models. Toman (1987) derives A-optimality results in a number of models, and returns to A-optimality in heterogeneity models in Toman and Notz (1991). Generally speaking, these articles solve the approximate design problem, although nearly all of these works address the issue of rounding to integer designs. Toman and Notz (1991) in particular give a new rounding strategy by rounding the amount (in the approximate theory) of the control in a treatment vs. control problem. They give some evidence that this rounding strategy gives better efficiencies than the methods suggested elsewhere.

6.3.3. D and E-optimality. D and E-optimal Bayes designs are discussed in Giovagnoli and Verdinelli (1983) and later in Toman and Notz (1991) for block models. In fact, Giovagnoli and Verdinelli (1983) show that in a two way ANOVA model, with usual as-

sumptions, there is a unique optimal design for fairly general criterion functions, and in the case of one treatment, a universally optimal design exists as well (universal optimality refers to simultaneous optimality under all criterion functions entertained). DasGupta and Studden (1991) also give instances of E-optimal designs in one way ANOVA settings and show that the E-optimal design coincides with an A-optimal solution.

6.3.4. Exact classical optimal designs. Although there is no direct evidence to this effect, results in the regression case and simple common sense suggest that exact classical optimal designs can be approximate proxies for Bayes solutions in most instances, if multivariate normal priors are used. In any event, if an exact classical optimal design for a problem is known, it can be instructive for the corresponding Bayes problem. Chapter 2 in Shah and Sinha (1989) has some general information on classical optimal designs, particularly some advantages of using balanced designs in block models.

The major bulk of the classical theory seems to have been for the D-optimality criterion. In particular, either algorithms or even catalogues of D-optimal designs for 2^m and 3^m factorial models are available for small values of m: one can see Nalimov (1982). The two most important contributions are the continuous D-optimum designs for second order models and central composite designs which proxy the theoretical continuous optimal designs. We describe these briefly below.

(a) Continuous D-optimal designs. Farrell, Kiefer and Walbran (1967) consider a general polynomial model in m factor experiments and describe the form of an optimal design depending on the degree of the polynomial and the number of factors, for three important design spaces: cubes, spheres, and simplexes. There is an interesting discussion on the minimum number of points that must be in the support of an optimal design, and using the methods of orthogonal arrays introduced by Rao (1946-1947), a device to reduce the number of points in the support of a candidate design is given. For those interested in the classic properties of designs, there is an intriguing example in which the optimal design is demonstrated to be not rotatable.

(b) Central composite designs. The conceptual difficulty with the theoretical continuous optimal designs in these cases is that they cannot be implemented due to their infinite support. One is thus forced into some form of an approximation. The by now

common method of approximation is to use what are called central composite designs. It is useful to know that central composite designs provide a rational way for moment matching corresponding to the continuous optimal designs; on the other hand, statistically they have a lot of simple appeal. To understand the structure of central composite designs, it is useful to think of the spherical design space once again. One can hit the boundary of the sphere by using points which have all but one coordinate equal to 0 and the remaining one equal to $\sqrt{m}$. Such points are called star points. Replications corresponding to 0 level for each factor are called center points. The remaining observations are used up in a 2^{p-f} fractional factorial for some $f > 0$.

Box and Hunter (1957) is a good exposition to central composite designs. Illustrative examples on the efficiencies of central composite designs according to the D-optimality criterion are available in Atkinson and Donev (1992).

These classic methods are of importance to the Bayesian theory at the present time for two reasons: first, the present state of the Bayes optimality theory is far less advanced than the classical theory. We just do not have very good knowledge of Bayes solutions yet. Second, it is always instructive to consider classic methods that are timetested and see if they do an adequate job when prior information is available.

7. Critique of the optimality theory.

7.1. Model dependence. A major criticism of the standard optimality theory in regression models is the severe dependence of the design on the assumed model. While it is a clear mathematical truth that if the model of a simple linear regression is valid, then nearly every criterion calls for taking observations only at the endpoints of the design interval, it is rare for even the most ardent theoretician to recommend this to anyone interested in analyzing data. This can be restated as saying that one never believes the model exactly. In this sense, despite the undisputable structure of the optimality theory and the impact it has had on further theory, optimal designs have not had much of an influence on people who do design actual experiments. Indeed, it is common in regression problems to more or less divide the observations uniformly in the interval. This has some similarity to D-optimal designs for high degree polynomial regression (this should not be interpreted as a weak convergence to the uniform distribution; indeed that is known to be not true). Even

Bayes designs, by virtue of sharing the same mathematics as their classical analogs, suffer from the same problem (in some problems, there are counterexamples to this; for instance, see the computer experiments in Mitchell, Sacks and Ylvisaker (1994)).

Another problem with the optimality theory is that the designs can be extremely goal specific; thus, in polynomial regression, a design which is optimal for estimating one regression coefficient may perform quite poorly for estimating another coefficient. The problem is that the experimenter may have several aims in an experiment, and in some problems other aims may arise at a later point of time (of course, it is not the design's fault that arbitrary aims specified afterwards cause problems). But the point remains that orientation to one specific goal is another drawback of the exact optimal designs.

7.2. Mixture models and model robustness. Response to the criticism of model dependence has been generally of three kinds:

(a) construction of designs that are nearly optimal in a lower order model with (good) protection against a higher order model,

(b) construction of minimax type designs in a class of models,

(c) use of criteria which are themselves some type of mixtures or averages of efficiencies under different models.

7.2.1. Protection for higher order models. Stigler (1971) proposed derivation of designs that are nearly G-optimal in a lower degree polynomial model with some protection for each model of higher degree upto a certain maximum degree (recall that G-optimality corresponds to an average c-optimality with an averaging over c: Studden (1977) calls this the integrated variance criterion). In Studden (1982), a D-optimal variant of Stigler's proposal is considered and designs which are most efficient at order r subject to a prespecified efficiency for an extra $m - r$ are derived. The results are remarkably closed form, with clever use of canonical moments of the design measure. As a matter of fact, tables of efficiencies are provided, and the general moral of this article is really quite encouraging: excellent efficiency at lower order models can be obtained by sacrificing a bit for the extra coefficients. It is also possible to show that the design supported on $\{0, \pm.618101, \pm 1\}$ with weights .13796, .140347 and .290673 is 69% $D-$ efficient for each of linear, quadratic,

cubic and quartic regression. This design maximizes the minimum D-efficiency over the polynomial models of degree $p = 1, 2, 3, 4$ over $[-1, 1]$; further information of theoretical nature is available in Dette and Studden (1994a). A similar formulation is the following: for a given design $\mathcal{E}$, consider estimating the mean response at value x of the independent variable when the degree of the polynomial is p. Suppose $v(x, \mathcal{E}, p)$ denotes the variance. Then consider the design that maximizes the minimum efficiency over $1 \leq p \leq n$ with respect to the integrated variance criterion $\int_{-\infty}^{\infty} v(x, \mathcal{E}, p)\ w(x)dx$ where $w(\cdot)$ is a probability density on the real line; this is the integrated variance criterion in Studden (1977). If $w(\cdot)$ is taken as the $N(0, 1)$ density and the design space $\mathcal{X}$ is taken as $[-1, 1]$, then the maximin design for $1 \leq p \leq 4$ is $\mathcal{E}(0) = .2296, \mathcal{E}(\pm.7018) = .2486$ and $\mathcal{E}(\pm 1) = .1366$. One can see plots of effficiencies of this design if it is used for particular x at the end of this article for each of $p = 1, 2, 3, 4$. A perception problem with these designs is that they are still equally thinly supported.

7.2.2. Neighborhood models. Huber (1975) suggested use of models in a neighborhood of a given model, and subsequent use of a minimax design, treating this as if this was a game with nature choosing a model from the neighborhood class. Of course, neighborhoods can be defined in many ways. Huber gives as an example the case of simple linear regression as a starting model and an $L(2)$ neighborhood of the linear regression function as the nature's family of models. It turns out that the $L(2)$ neighborhood does not quite work very well, but clearly the suggestion in Huber (1975) is appealing.

The idea in Huber (1975) was picked up again in Marcus and Sacks (1976), where they look at $L(\infty)$ type neighborhoods of a linear regression function and under various envelopes for the family, derive minimax (estimate-design) pairs. But again, the designs are thinly supported. They find that as a rule, use of an estimate with a nonoptimal design is more dangerous than use of an optimal design with a "nonoptimal" estimate. This leads to some qualitative understanding.

The ideas in both Huber (1975) and Marcus and Sacks (1976) need further attention. Another recent article is Tang (1993).

7.2.3. Mixture criteria. The main idea is to take an indexed family of functionals (ϕ_p) and then use an average of these functionals over p. The averaging is done by using a

subjective weight measure on p. There are two leading articles on this approach; Dette and Studden (1994a) give a very elegant theory using D-optimality as the basic criterion and p as the degree in a polynomial regression. In Dette (1992) similar kinds of results are derived for polynomial regressions with missing powers. Elegant use of canonical moments and continued fractions can be seen in these articles.

7.3. All around designs. As stated earlier, the optimality theory of experimental design also suffers from orientation towards very specific goals. For instance, in quadratic regression on $[-1, 1]$, the optimal design for estimating the coefficient of x^2 has efficiency .5 for estimating the intercept. It is thus an useful exercise to investigate if designs can be found which give good efficiencies simultaneously for a number of aims. As commented before, it is necessary that these aims be stated before the experiment as addition of new goals afterwards essentially makes the problem impossible. Since all the aims of an experiment are impossible to treat or even conceive of in a theoretical study, an effort to find all around designs has to be quite specific. Thus a theoretical study of this issue is limited in its scope, but an investigation in some standard models with standard goals can lead to an appreciation of the extent to which all around designs are plausible.

Lee (1988) discusses this under the terminology of constrained designs. Under the general structure of minimizing $\phi_{m+1}(M)$ subject to $\phi_i(M) \leq c_i, i = 1, 2, ..., m$, for differentiable $\{\phi_i\}$, Lee gives an equivalence theorem and in particular, cites as example a D-optimal design in quadratic regression with an upper bound on the trace of the dispersion matrix. Essentially the same approach is seen under less smoothness assumptions on the functionals $\{\phi_i\}$ in Pukelsheim (1993).

DasGupta, Studden and Mukhopadhyay (1992) take the above functionals to be the variances of the leastsquares estimates of individual coefficients in a general linear model with a general variance function and derive designs that have a guaranteed efficiency e for each coefficient. The largest e for which such a design exists is of interest. For the corresponding Bayesian problem, they substitute posterior variance of the parameter for variance of the estimate. Two special variance functions are subsequently used. An interesting fact is that good efficiencies can be guaranteed if the subscripts of the coefficients of interest are all even or all odd. The following is an illustration in the case of cubic

polynomial regression:

Subscripts of coefficients	Guaranteed largest efficiency
0, 2	.75
1, 2	.65
1, 3	.93
0, 1, 3	.66
0, 1, 2	.58
0, 1, 2, 3	.58

8. Nonconjugate priors and robust Bayes designs.

8.1. Nonconjugate priors. A major problem in using nonconjugate priors in the canonical normal linear model is the associated loss of closed form formulae for the Bayes risk; at a more fundamental level, unlike in the case of normal priors, now the posterior expected loss in fact does depend on the actual observations that would be later obtained. Thus posterior expected losses can no longer be used for minimization, and even Bayes risks do not have closed form expressions. Some work is going on at the present time at Purdue University on Bayes design with nonnormal priors. A combination of some theory and subsequent computations, this work essentially follows the following line:

(a) One proves that the Bayes risk for estimation of the coefficients of the model is decreasing and convex in the information matrix in the Loewner ordering;

(b) For general regression functions, one uses a Caratheodory bound and for special types one uses the Kiefer bound on the number of support points;

(c) One then uses the Brown-Stein (Brown (1986)) identity for Bayes risk, but now specialization to (a general) quadratic loss (as in Owen (1970)) is necessary;

(d) One finally does a numerical search for the optimal design.

This method has nothing to do with conjugacy or otherwise of the prior. There is some dimensionality reduction for symmetric priors if the design space is symmetric (i.e., $x \in \mathcal{X}$

implies $-x \in \mathcal{X}$). One can start with other criteria functionals that already satisfy property (a) and the same steps then follow through. The value of using nonconjugate priors stems from two directions: *i.* in some problems, the experimenter does not want to use conjugate priors, and ii. as a general intellectual question, it is necessary to know if conjugate vs. nonconjugate priors really do result in significantly different Bayes solutions for small sample sizes. For sample sizes that are large compared to the number of parameters, we do not believe conjugacy vs. nonconjugacy matters. We do not have at the present time a clean prescription for what constitutes a large enough sample size for a given number of parameters, but one should consult Berger (1986) for further discussion on this issue. The gist is that inference and design seem to behave as fundamentally different problems with respect to fine specification of the prior if frequentist Bayes design criteria are used. We do not have any knowledge, however, if this is the case if preposterior design criteria are used: preposterior criteria would be discussed at length in section 10.

8.2. Robust Bayes designs.

There is now a substantial literature on robustness of Bayes methods, to various components in a decision theory framework, although a majority of the work is on robustness with respect to the prior. Berger (1994) and Wasserman (1992) give a lot of comprehensive information and food for thought. All robustness work, frequentist or Bayes, fall into one of two general categories: *i.* take a fixed procedure optimal under one model and ask what it does for another (close) model, and *ii*). take a family of models close together and ask what procedure(s) provide protection for all these models. There is a third way to look at robustness: what is needed for robustness to (honestly) obtain; for example, is symmetry essential, or is an exponential or faster tail essential, does one need a good idea of the variance, etc. There does not seem to be much work on this view of robustness, perhaps because there is a belief at large that NOTHING is needed and more thinking or more data would solve the problem. We recommend chapter 1 in Huber (1981) and Staudte and Sheather (1990) and Rubin (1977) for anyone wanting to learn about robustness.

8.2.1. Regression models. DasGupta and Studden (1991) consider the normal linear model with a family of priors and take the following general approach: pick a special prior from the family; pick a criterion functional; then derive a design that gives the best robustness for the family of priors subject to being ϵ-optimal with respect to the special prior. The

article then considers various criterion functionals, various inference problems, two different families of priors, and describes the necessary mathematics to solve these problems. The two families of priors were respectively first suggested in the Bayesian robustness literature by Leamer (1978) and Polasek (1985), and DeRobertis and Hartigan (1981). The necessary mathematics mostly is convexity-invariance with some moment theory for the DeRobertis-Hartigan family, a density band for the prior density. A companion article DasGupta and Studden (1988) also has material directly relevant to Bayes designs that in addition uses geometric argument. The following is one result from DasGupta and Studden (1991):

Theorem. Consider a density band for the prior with envelopes that are multiples of a given normal density; then the design that gives the smallest confidence set with a given posterior probability for each prior in the density band is the Bayes D-optimal design with respect to that given normal prior density.

One can see a variety of other results in that article.

8.2.2. Other models. There is some literature on robust Bayes designs for other special problems; almost all of this work is due to Blaza Toman and her coauthors. Toman (1992) and Toman and Gastwirth (1993) are two important articles on ANOVA models in the context of robust Bayes designs. The most important thing about both of these articles is that the optimization is simultaneously over the (estimate,design) pair. There is something to be said for this viewpoint, and therefore these two articles contribute a formulational novelty in this area. In Toman (1992), the priors are essentially the Leamer-Polasek type, while in Toman and Gastwirth (1993), they are finite mixtures of normals. In each article, the ultimate criterion is an average over the family of priors, with some difference in the details. For instance, Toman (1992) has an information theoretic criterion. The results are basically closed form in both articles.

If one takes the view that Bayesian statistics should be robust subjective, then clearly much further work remains to be done. However, for sample sizes that are not very small, one is likely to see little sensitivity to the prior because the Bayes solutions for different priors would all be close to the classical solution and therefore close to one another. This is not a precise statement, but only underscores the qualitative phenomenon.

As a historical point, robust Bayesians should also see Huber (1972) and chapter 10

in Huber (1981).

9. Miscellaneous other design problems.

9.1. Quality control.

Since the late seventies, there has been a change in attitude about what process quality control really means. While acceptance sampling and $k\sigma$ control limits dominated the practice of quality control for a very long time, new thoughts emerged in the late seventies. Thus, although the traditional methods of tolerance specification and acceptance sampling still have some role in classroom teaching and actual process control in the US, offline production control appears to have taken over in other parts of the globe, in particular Japan. We recommend the articles Kackar (1985) and Pukelsheim (1988) for excellently presented expositions on this topic. Also see Ghosh (1990).

The idea of offline production control is the design or adaptation of the process parameters such that the target is attained more or less exactly, and variance is minimized subject to target attainment. This is in contrast to the traditional method of estimating process capability vis-a-vis tolerance limits. A short but pertinent article on Bayes design with normally distributed observations with normally distributed parameters in the context of offline process control is Verdinelli and Wynn (1988). One should also see Sarkadi and Vincze (1974) for a systematic presentation of mathematical problems in quality control.

9.2. Engineering design and reliability.

There is a fairly substantial amount of work on Bayes design relevant to reliability, which is comprehensively covered in Chaloner and Verdinelli (1994). We are not aware of any kind of a Bayes optimality theory in Engineering design problems; however, the monograph of Wilde (1978) followed by Papalambros and Wilde (1988) give well written introductions to a wealth of really interesting problems.

9.3. Spatial designs.

Generally speaking, spatial design corresponds to problems of prediction or otherwise of a response when the input variable is spatial. Thus the design set may be a finite set of points with three coordinates each, identifying each point with the geographic location of an experimental station. Problems in multivariate numerical analysis such as approximation of an integral or evaluation of other linear functionals by averaging over a discrete set of

points also go by the general name of spatial optimal designs.

There are two intrinsic features that stand out in these problems:

i. depending on the exact criterion used, the fundamental nature of the design such as thickness of the support and high density regions can change and

ii. significantly more than standard problems, computation of the optimal design becomes an issue. In fact, it seems that development of computing algorithms may even be the most dominant issue.

The general tendency in these works seems to have been to take the viewpoint that the response behaves like the path of a Gaussian process with spatial time variables. The experimenter's belief about the smoothness of the response is incorporated into the autocorrelation structure of the Gaussian process. We recommend Sacks and Ylvisaker (1970), Diaconis (1987) and Sacks and Schiller (1988) specifically for reading on spatial designs. Diaconis (1987) in addition takes the reader along a fascinating path on the history of intellectual efforts to connect probability and statistics with numerical analysis problems.

10. Conditional formulations and sample size choice.

10.1. Conditional formulations.

Traditionally, experimental design is regarded as one area in which even the strict believer in conditional Bayes has to resort to an integration on the sample space. The reason is obvious: at the design stage, there are no data, and so design criteria necessarily have to average over potential data that may arise, which corresponds to frequentist integrations on the sample space. Thus, for instance, Bayesian A-optimality calls for minimizing the Bayes risk, a double integral on the joint probability space.

However, conditional or preposterior formulations of the design problem are possible, although even this formulation has a frequentist flavor. In other words, even in the preposterior formulation, considerations of the totality of the samples cannot be completely ignored. DasGupta and Vidakovic (1994) and DasGupta and Mukhopadhyay (1994) give detailed discussions of this; we will give a sketch of the preposterior formulation in these articles.

Consider the canonical normal linear model, with some prior on the parameters. Suppose we accept posterior A-optimality as the criterion, again for specificity. Given a design, and once the actual data arrive, there is a posterior density for the parameters. The desirable goal is to minimize the trace of the posterior covariance matrix. However, this cannot be done since in general the posterior covariance matrix certainly depends on the data. In the preposterior formulation, one specifies an upper bound on the trace of the posterior covariance matrix and chooses a design that satisfies this upper bound, with a prespecified large predictive probability. In other words, for data that are likely to arise, the design already gives a desired accuracy. For very ambitious upper bounds on the trace, no such design may exist, but the mathematics of the problem will say so if that is the case. In certain cases, the upper bound can be satisfied for all possible samples, i.e., with a predictive probability of 1. Naturally, the predictive probability is with respect to the predictive distribution conceived from the given prior. Alternatively, this preposterior formulation can be written in a minimax type of statement:

$$Min \;\; Max \;\; tr(V(y|M,\, \pi)),$$

where the maximum is over a set of samples y with predictive probability $1 - \epsilon$, the minimum is over the information matrix (i.e., design) M, and $V(y|M,\ \pi)$ denotes the posterior covariance matrix for given y and M. Note that a SPECIFIC set of predictive probability $1 - \epsilon$ has to be used; but this specific choice should usually be an obvious natural choice (for instance, a high density set of the predictive distribution).

There has not been any work with the preposterior formulation other than for Bayesian sample size choice, which we consider next.

10.2. Minimum sample size.

10.2.1. The formulation. The general formulation of the minimum sample size problem is the following: one specifies a measure of accuracy for the particular inference problem at hand, and asks what is the minimum sample size n for which the accuracy requirement is met. For instance, for estimating a binomial proportion, it is standard to use the expected length of a 95% confidence interval and seek a sample size that makes it smaller than a given upper bound. Bayesian considerations automatically enter into such classical calculations, because the expected length is a function of the unknown proportion and

therefore an apriori guess value needs to be used to arrive at a sample size not overly conservative. As another example, in testing problems, it is standard to seek a sample size that ensures a given power for some standard 5% test at a value of the parameter thought to be practically different from the null value. One can dispute the correctness of these formulations; but we are only making the point that this is how it is traditionally done in classical statistics. There is absolutely no doubt that minimum sample sizes are taken seriously by people who deal with data, and this topic has assumed the status of textbook material in traditional statistics. Students are told about it. There are a large number of monographs, books, tables and charts of classical minimum sample sizes. We recommend Odeh (1975) and Selected Tables in Mathematical Statistics (1975) as two particular sources for further exposition and actual numbers for possible use.

The corresponding Bayesian formulation can be of two possible types: a preposterior formulation of exactly the kind we described above, and a frequentist Bayes formulation in which one seeks a sample size that ensures that the Bayes risk in the problem at hand is smaller than a given number. The preposterior formulation is more Bayesian in the sense of not integrating on the sample space and is also by leaps and bounds mathematically more challenging. However, this itself can be a negative aspect of the preposterior formulation, in which case the frequentist Bayes formulation can be adopted. We must admit,however, that the frequentist Bayes formulation can lead to a completely trivial problem, though not always.

10.2.2. Normal theory. A large spectrum of normal (univariate and multivariate) problems with the preposterior formulation and associated theory and in some cases actual sample sizes are available in DasGupta and Mukhopadhyay (1994) and DasGupta and Vidakovic (1994). In DasGupta and Mukhopadhyay (1994), a theory of Bayesian sample sizes is provided for two problems: *i.* testing for a normal mean, with conjugate priors, and seeking a sample size that either makes the posterior risk uniformly small or makes the posterior risk uniformly robust if one takes a family of priors instead of one specific prior. There are nontrivial asymptotics in these problems; in this same article, DasGupta and Mukhopadhyay (1994) also give some actual Bayes sample sizes, although their practical adoption in the forseeable future is at least doubtful; *ii.* constructing a confidence set for a multivariate normal mean, for conjugate priors, and seeking a sample size that gives a set

with a prespecified posterior probability uniformly over future data and simultaneously for every prior in the family under consideration. Table 1 in the same article gives some actual sample sizes for this as well. A previous longer version DasGupta and Mukhopadhyay (1992) had a number of other problems in which the preposterior formulation was discussed and a theory presented; among the other problems in this longer version was the nonregular case where the sample space depends on the parameter and there is also some consideration of an uncertain loss function.

The widely used one way ANOVA model is considered in DasGupta and Vidakovic (1994). The problem is testing for no treatment differences, and the approach is purely Bayesian. That is, a prior probability is given for the hypothesis, and a normal prior density used as the conditional density of the parameters given that the null hypothesis is not true. Then testing is regarded as a decision problem with 0-1 loss, i.e., the quantity they try to keep small is the posterior probability of the wrong hypothesis being picked. Again, some actual sample sizes are given, but now a complete Mathematica code is provided for use by the specific user with his/her particular inputs.

10.2.3. Binomial proportions. The approach taken for determination of Bayesian sample sizes in this case has generally been the interval estimation approach. That is, one seeks a sample size that ensures a posterior confidence interval of a specified probability such that its length is smaller than a given number. Again, the length is a function of the data, and either the expected length (under the predictive distribution) or the maximum length (over all possible data) are substituted for the actual length. A conceptual difficulty with the expected length is that there is no guarantee at all that by using the sample size produced by this criterion, the accuracy goal one started out with would be satisfied when data do arrive. One then feels the exercise was useless. Therefore, although more conservative, the maximum length criterion is preferable. So far the work has assumed conjugate Beta priors. This is just fine, because in this case, Beta priors can approximate any prior whatsoever on $[0, 1]$ by simply allowing mixtures. So any generalization, if at all, that is needed is consideration of some Beta mixtures. The work on Bayesian sample sizes for Binomial proportions is due to Lawrence Joseph and his coauthors; one should in particular see Joseph, Wolfson and Berger (1994) and Joseph and Berger (1994) and an earlier work Bock and Toutenberg (1991) in the context of clinical trials.

10.3. Optimal sample sizes.

10.3.1. Cost vs. accuracy. Statistical theory often has the pretence of being able to choose as many samples as it takes to ensure a given accuracy. This of course is far from the truth in many real problems. The impediments are many: cost of sampling is the most prohibitive factor; in some situations, sampling may be a time consuming process or simply a difficult human exercise. Optimal sample sizes try to balance the tradeoff between accuracy of inference which (generally) increases with increasing sample size and sampling and other human costs which also increase with increasing sample size. Theoretically, it is assumed that all of these costs are considered together in a single cost function although quantifying extraneous costs like the value of time does not appear to be an easy task. The optimal sample size is only to be taken as a guideline; it is certainly not intended as a rigid prescription. There are also some conceptual debates whether abstract (and often unitless) quantities such as inference accuracy and dollar amounts for cost of sampling can be just added to form the overall cost; one can see Chernoff (1972) for a discussion of this.

10.3.2. Estimation. Suppose one has a loss function for accuracy, $L(\theta,\ a)$, and a cost function $C(n)$; then an overall cost is the sum total of the two. Suppose it is decided to use a rule $d(X)$, which should be just the Bayes rule for the given loss function and a given prior in a Bayesian framework, and would be some classical estimate like an mle or a best invariant rule otherwise. Then the total average risk is $R(\theta,\ d(X)) + C(n)$, where θ denotes the generic parameter. Minimizing this with respect to n would entail a solution that depends on (part of) the unknown parameter. It is therefore natural to take the Bayes risk in place of the risk function $R(\theta,\ d(X))$ and minimize the resultant quantity with respect to n. If no prior distribution is used, one can use a guess value for the parameters present in the solution or alternatively use sequential versions which estimate the parameters at each stage, use the estimated parameters in the formula for the sample size, and use the first n satisfying appropriate constraints. Details of such an approach can be seen in Starr and Woodroofe (1969), Ghosh, Sinha and Mukhopadhyay (1976) etc.

10.3.3. Testing. The steps in determination of an optimal sample size are the same in any decision problem; however, for testing problems, the loss functions associated with the

two actions when there are two hypotheses are different from standard losses one sees in estimation. 0-1 types of losses where there is no loss in accepting the correct hypothesis and a constant loss in accepting the wrong hypothesis are probably the most used, although appear to be severely unrealistic. If by chance or due to any other reason, a seriously wrong hypothesis is accepted, the penalty should be more than accepting a hypothesis which is false only on paper. For instance, for deciding the sign of the mean of a univariate normal distribution, the loss in deciding the wrong sign could be reasonably taken as a nonconstant monotone nondecreasing function of the absolute value of the mean.

Berger (1986) and Chernoff (1972) both give a nice normal theory example by taking the absolute value of the mean as the loss for inferring the wrong sign and derive a Bayesian optimal sample size using a conjugate prior. Chernoff (1972) also does the case of a simple vs simple hypothesis and shows the qualitative difference between simple and nonseparated hypotheses. One should also see Antelman (1965), a charming article, which has much to offer to people interested in Bayes experimental designs as a whole.

11. Nonlinear problems.

11.1. Introduction.

Nonlinear problems can arise either for nonlinear functions in a linear model, or in models where the response function is itself nonlinear; these latter class of models is more or less universally known as nonlinear models. We recommend the beautiful yet encyclopaedic book due to Seber and Wild (1989) to anyone interested in nonlinear models. Of particular value for workers interested in experimental design are chapters 1, 5, 6, 7, 8, 9 and 10; section 5.13 gives a short but excellent historical account of optimal design theory in nonlinear models, complete with early examples from Box and Lucas (1959), early criteria, and discussions about thin designs. Our discussion of nonlinear functions and nonlinear models would be short due to a second article on nonlinear models. In addition, Chaloner and Verdinelli (1994) give a fairly complete account of the current bibliography on Bayes optimal design for nonlinear models.

11.2. Nonlinear functions in linear models.

An early example of an interesting nonlinear function is the example of calibration

in simple linear regression: estimating the value of the indepenent variable at which the mean response is zero or some other constant; this is usually known as Fieller's problem. Also see Rao (1973) for a clever description of how to construct confidence intervals in this seemingly impossible problem. Silvey (1980) gives an elementary but very insightful account of optimal design for this problem. Another example essentially the same as Fieller's is estimating the value of the independent variable at which a quadratic response function has a zero derivative. This is commonly known as the turning point problem.

The major conceptual problem with experimental design in nonlinear problems is that the optimal design depends on the very parameters one is trying to learn about. It has been suggested that one uses a guess for the value(s) of the parametric functions that enter into the design, much as it is customary in most elementary texts to use a guess for the value of a Binomial proportion in order to construct sample sizes for learning about the same proportion. This is also conceptually akin to construction of locally most powerful tests where one maximizes the derivative of the power at the null value; see Lehmann (1986). This approach to design has been called local optimality; see Chernoff (1972) and Atkinson and Donev (1992). The locally optimal design is exactly optimal if the true value of the parameter happens to be equal to the guess value. However, use of locally optimal designs can subsequently lead to amusing problems: see Chaloner and Verdinelli (1994).

Other examples of nonlinear functions in linear models include estimating the probability that a future observation belongs to a specified set, say a bounded interval; other linear models in which standard problems cause nonlinear functions to arise are models in which the variance is a function of the mean. One should also see Wu (1988) for treatment of optimality theory for estimation of similar nonlinear functions in quantal response models.

11.3. Nonlinear models.

11.3.1. Basic tool. Any statistical model in which the mean response $E(Y)$ is a general function $f(\underset{\sim}{\theta}, x_1, x_2, \ldots, x_p)$ of p independent variables and a (possibly vector valued) parameter $\underset{\sim}{\theta}$, is a nonlinear model. The goals of experimentation in such models may vary, as in linear models. A persistent common feature of optimal experimental designs across these models is that the design depends on the unknown parameter; some isolated

counterexamples to this statement are known - in particular, one can see Silvey (1980).

A key tool for the optimality theory in nonlinear models is a general equivalence theorem in Whittle(1973). As is the case with any equivalence theorem, the optimality of a candidate design can be either disproved unambiguously or established subject to the accuracy of the computations by using Whittle's theorem. This is because any equivalence theorem is a statement of the following kind: a design $\mathcal{E}$ is optimal if and only if an appropriate functional $F(\mathcal{E}, x) \leq 0$ for every x in the design space and is $= 0$ for exactly those points in the support of $\mathcal{E}$. Now, by using only calculus and numerical optimization, it is usually feasible to find the best design with k points in its support, for a given k. This candidate design can then be tested for optimality by using the equivalence theorem. The difficulty might be that for points in the support of this candidate design, the functional F may give values which are small negative numbers. One is then forced to make a judgement if this is only an accuracy problem or the design is not the exact optimal one. In any event, such an use of the Whittle equivalence theorem is clever and was initiated by Chaloner and her coauthors : one can see in particular Chaloner and Larntz(1986) and Chaloner(1993).

11.3.2. State of the art. The current inclination in the choice of a criterion for optimization seems to be to take an information approach, most likely influenced by Lindley (1956). Typically, one takes the Fisher information matrix $I(\underset{\sim}{\theta}, \mathcal{E})$ for a given design $\mathcal{E}$ and takes the logarithm of its determinant as the starting step: the logarithm just happens to cause a great amount of algebraic simplicity in many problems, and is therefore justified on the basis of the rewards it produces. But one still has the parameters in the criterion and therefore it seems natural to take an average of it with respect to a weight function, which may or may not be a prior, but operationally acts like a prior. As a matter of fact, this is the only place where the prior has anything to do with the problem and therefore even those averse to use of priors can pragmatically adopt this approach if local optimality is not regarded favorably. One can see Chaloner and Verdinelli (1994) for some additional discussion on use of the Fisher information as a basis for optimality.

A disappointing but apparently unavoidable feature of the optimality theory in nonlinear models is that general complete class theorems about admissible designs in terms of the number of points in their support seem very difficult, and probably impossible. However,

some fairly unexpected advances have been made in the last few years; foremost among these surprising and hard works are Dette and Neugebauer (1993) and Dette and Sperlich (1994a, b). The typical tone of these papers is the following: given a prior distribution on the parameters, they characterize situations when an optimal design with a given number of points in the support exists, and if one does, identify the points and the weights more precisely. There is also an attempt to establish analogs of Caratheodory type bounds on the cardinality of the support. Despite the fact that these advances are model specific, they are distinctively strong results. Earlier, Mukhopadhyay and Haines (1993) also took similar approaches in an exponential model.

It is very important to be aware of a fundamental phenomenon that is emerging as a unified character of the optimality theory in nonlinear models: an opinionated prior results in an optimal design similar to the corresponding locally optimum design, and a flat prior results in thickly supported designs. However, as of this time, the optimality theory still suffers from the same drawback as for linear models: it produces designs that are unlikely to be adopted in practice.

11.3.3. Linearization of a nonlinear model. Atkinson and Donev (1992) describe a method for linearizing a nonlinear model by using a Taylor series expansion, thereby producing polynomial models, and iteratively fitting the model in the approximate form. The general idea of linearizing a nonlinear model seems to have some potential; however, there are subtle points in using such a method. One possibility is to find a uniform approximation to a given degree of accuracy for the true response function by using response functions we know how to handle, say polynomials, find an optimal design in the linearized model, estimate the parameters in the linearized model, and finally transform these estimates back to the initial model: THIS LAST STEP IS SUBTLE, because the parameters that appear as coefficients in the linearized model would not be the parameters for which a prior was elicited, but functions of those parameters. So the retransforming may have to be done by an inverse function method, or a pure Bayes method, though the pure Bayes method is the harder one to use. As for linearizing a function to produce uniform approximations in a compact design set, many methods are available. Expansions in Chebyshev polynomials, exchange algorithms and others are common tools with theoretical properties; one can see Powell (1981) and Rivlin (1990). We will return to this topic from a technical angle in

section 13(appendix).

12. Future of Bayes design.

It seems as though the future of Bayes experimental designs lies in nonstandard problems, as opposed to standard linear models or even standard nonlinear models. Bayes optimality theory can succeed as a useful theoretical development only if it is seen that the resultant theory does not reproduce the classical solutions either exactly or practically exactly. We have seen a few examples of Bayes optimal designs which differ remarkably in their character with usual optimal designs which are embarrassingly thinly supported: instances of these are Mitchell, Sacks and Ylvisaker (1994) and Sacks and Schiller (1988).

There are innumerable interesting problems in various branches of science where optimal design is a very viable scientific issue; it seems imperative that subjective prior information is available in many of these problems, and thus Bayes design should have a useful role to play. It may turn out that the Bayes solution would once again track classical methods closely or exactly. We believe such findings, although negative in a sense, would be intellectually valuable.

Examples of areas where optimal designs can be explored and promise to be interesting scientific exercises are indeed numerous: inventory control, tracking a moving target, design of neural networks, random variate generation, structural and engineering design, construction of histograms, survey sampling, combinatorial algorithms such as the travelling salesman problem, clustering, markov decision processes, and so on. We are aware of some work relating to design in progress on a few of these areas.

13. Appendix.

13.1. Moment methods.

13.1.1. Markov moment problem and its geometry. Moment theory enters into the considerations of optimal design through the concept of admissible designs, as discussed in section 6. The problem there is to find a probability distribution that maximizes the $2pth$ moment for given values of the preceding moments.

Example. Suppose one is interested in finding the maximum value of the fourth moment of a distribution on $[-1, 1]$ with the first and the third moment equal to zero and the

second moment equal to some given c.

The idea then is to construct a polynomial of degree 4 of the form

$$P(x) \;=\; x^2\,(x-1)\,(x+1),$$

which has the property that $P(x) \leq 0$ for every x in $[-1,1]$ and is equal to zero if $x = 0, -1$ or $+1$. Thus, for any probability distribution, the fourth moment c_4 satisfies $c_4 \leq c$, and for the particular distribution which assigns probability $c/2$ at $+1$ and -1 and $1-c$ at zero, equality obtains because $E(P(x)) = 0$ if the underlying distribution is supported on the roots of $P(x)$; alternatively, in this simple case, equality can as well be seen by trivial direct verification.

Karlin and Studden (1966) and Kemperman (1968) give very careful accounts of such geometry of moment problems; the general version of a moment problem, which sometimes goes by the name of the Markov moment problem, is the following:

One has a general set, on which are defined $k+1$ functions $f_0, f_1, \ldots, f_k$. The problem is to determine a distribution $\overline{\mathcal{E}}$ and $\underline{\mathcal{E}}$ that respectively maximizes and minimizes the integral $\int f_k d\mathcal{E}$ among all distributions satisfying $\int f_i d\mathcal{E} = c_i, 0 \leq i \leq k-1$.

Example. Suppose X has a density on R of the form

$$f(x|\mu) \;=\; \int \frac{1}{\sqrt{2\pi s}}\, e^{-\frac{1}{2s}(x-\mu)^2} dG(s).$$

In other words, the density of X is a normal scale mixture. Consider the interval $[x - 1.96, x + 1.96)$, which as a confidence interval for μ, has a 95% confidence coefficient for normal data. The object is to determine the smallest confidence coefficient of this interval if the underlying distribution has standard deviation 1.

It is immediate that one thus wants to

$$\text{minimize} \int \Phi(\frac{1.96}{\sqrt{s}}) dG(s) \text{ subject to } \int s dG(s) = 1.$$

This is therefore a special kind of a Markov moment problem.

To find a solution, one can show that there exists a straight line with equation $y = a + bs$, such that this line always lies below the graph of the function $\Phi(\frac{1.96}{\sqrt{s}})$ for $s \geq 0$ and

touches the graph at $s = 0$ and another point $s = s_0 \geq 1$. If one now takes a distribution G_0 supported on $\{0, s_0\}$ such that it indeed satisfies the constraint $\int s dG_0(s) = 1$, then, by virtue of the geometric property of the above straight line, it indeed follows that for any distribution with the stated restriction, $\int \Phi(\frac{1.96}{\sqrt{s}}) dG(s) \geq a + b$, and in particular for the distribution G_0, it is equal to $a + b$, because at the points of support of G_0, the linear function $a + bs$ and the function $\Phi(\frac{1.96}{\sqrt{s}})$ are exactly equal!

The value of s_0, and the constants a, b can be found by easy geometric considerations, and are all given in Basu and DasGupta (1994).

The technique given in these two above examples forms a key step in solution of moment problems: identify linear combinations $\sum_{i=0}^{k-1} a_i f_i$ which either lie below or above the function f_k, depending on whether the Markov moment problem is one of minimization or maximization, and find a distribution which is supported on the points at which the graphs of these two come into contact, and appropriately adjust the weights so as to satisfy the given moment constraints. Notice that the task involves finding just the right linear combination with the given geometric property; one should not be misled that there is one such linear function only.

Example. Does there exist a unimodal distribution on $[0, 1]$ with variance equal to .2?

The question is meaningful, because one can attain a variance between 0 and .25 if all distributions on $[0, 1]$ are allowed. To answer this question, it is helpful to turn it into a moment problem on writing the underlying unimodal random variable X as $X = a + UZ$, where the mode $'a'$ is between 0 and 1, U is uniformly distributed on $[0, 1]$, Z is independent of U, and is between $-a$ and $1 - a$ with probability 1. Of course, the mode $'a'$ is not fixed, and has to be varied between 0 and 1 as well.

It is clear that marginally, any mean between 0 and 1 can arise from a unimodal distribution; furthermore, all point masses are unimodal, and therefore the lower boundary of the relevant moment set for unimodal distributions is given by $\underline{\mu_2} = \mu_1^2$.

The upper boundary can be found easily after proving the following fact: a given mean μ_1 can arise only if the mode $'a'$ satisfies: $max\ (0, 2\mu_1 - 1) \leq a \leq min\ (2\mu_1, 1)$. Indeed, the upper boundary of the moment set is piecewise linear, given by $\overline{\mu}_2 = \frac{2}{3}\mu_1$ if $\mu_1 \leq 1/2$,

and $\frac{4}{3}\mu_1 - \frac{1}{3}$ if $\mu_1 > 1/2$. From this it follows by calculus that the maximum variance is $1/9$, which is much less than .2. Thus, no variance larger than $1/9$ can be attained by a unimodal distribution.

Notice that the geometry shows that the moment set for unimodal distribution is not convex.

We recommend Krein and Nudelman (1973) and Akhiezer (1965) for anyone interested in learning about moment methods, a very useful tool in many branches of mathematics and mathematical statistics. We also recommend Diaconis (1987) for a very lively account of the history of moment methods and also some simply interesting facts of inherent scientific interest: for example, suppose we have a CDF F on the real line which has exactly the same first n moments as those of a $N(0,1)$ distribution; how accurately does this determine the CDF itself?

Characterizing distributions which are determined by their moment sequences is also a celebrated problem in the history of moment theory and probability. The normal distribution is determined by its moment sequence, but alas, the lognormal distribution is not! Convolutions of determined distributions may be undetermined - see Berg (1985). On the positive side, any distribution which is boundedly supported is determined by its moment sequence. In fact, there is a peculiar generalization to this which really is a result from analytic function theory stated in the language of a probabilist.

Theorem. Let $\{n_i\}$ be a subsequence of the positive integers such that $\sum \frac{1}{n_i} = \infty$; then the sequence of moments $E(X^{n_i})$ determines a distribution supported on a bounded interval $[a,b]$.

There are indeed characterization theorems which, in theory, can tell which distributions are determined and which are not by their moment sequences. They are not particularly useful in general; there is a remarkable exception to this, a pretty theorem:

Theorem. Suppose an absolutely continuous distribution has a density $f(x)$ on $\mathbb{R}$. Then it is determined by its moment sequence if and only if $\int_a^b \frac{log\ f(x)}{1+x^2}\ dx = -\infty$.

This theorem has a counterpart for measures supported on $\mathbb{R}^+$ in the following sense:

Theorem. Suppose an absolutely continuous distribution supported on $\mathbb{R}^+$ has a density $f(x)$ such that $\int_0^\infty \frac{\log f(x)}{\sqrt{x}(1+x)} dx > -\infty$. Then there is at least one more distribution supported on $\mathbb{R}^+$ with the same moment sequence as that of f.

Although it does not appear to be well known, there is a striking connection of this result to Hardy functions: f is undetermined if and only if it is the absolute value of the Fourier transform of an L_2 function vanishing in the negative half line. A recent description can be seen in Berg (1995).

Whether or not a particular distribution is determined by its moment sequence can sometimes be useful in asymptotic theory; in general, convergence of all moments does not imply convergence in distribution. However, if the suspected limit distribution is determined by its moments, then convergence of moments does indeed imply convergence in distribution. One can see Billingsley (1986) for more on this; in particular, certain astounding results in probabilistic number theory which otherwise require very intricate sieve and truncation arguments, can be proved by moment convergence and by using the fact that the normal distribution is indeed determined by its moment sequence. If one is interested, the sieve arguments can be seen in Elliott (1979), who describes the proof of limiting normality of the number of factors of a random integer.

13.1.2. Canonical moments. In general, the first n moments of a probability distribution on a bounded interval $[a, b]$ satisfy complex inequalities; more precisely, if one defines a set in R^n as

$$M_n = \{(c_1, c_2, \ldots, c_n) : \ c_i = E(X^i) \text{ for some probability distribution on } [a, b]\},$$

then M_n is a complicated convex set in R^n. It usually goes by the name of the moment set.

Canonical moments are a device for transforming the moment set into the cube $[0, 1]^n$; thus, each canonical moment p_i varies freely in the interval $[0, 1]$ as opposed to the moments c_i which form a complex set. This transformation from moments to canonical moments is $1 - 1$ onto. Thus, it is quite common in optimal design theory to work out an optimal design in terms of its canonical moments, and the added bonus is that the structure of the optimizing canonical moment sequence even gives the number of points in the support of

the optimal design. One can see numerous evidence of the utility of this technique in the works of William Studden. Canonical moments are also obviously useful in any numerical optimization scheme over the moments, because the optimization can be done for freely varying variables, which cannot be done with the moments themselves.

The exact transform to take moments to canonical moments and vice versa is best described by a recursive algorithm; we recommend Skibinsky (1968) for this.

Example. Consider the Uniform distribution on [0,1]. The moments of this distribution are given by $c_i = \frac{1}{i+1}$; the canonical moments are seen to be $p_{2i-1} = 1/2, p_{2i} = i/(2i+1)$.

Example. Suppose $X \sim Bin(n,p)$; the moments of X/n can be calculated by calculating the factorial moments, $E\{X(X-1)\dots(X-i+1)\}$. The canonical moments are seen to be $p_{2i-1} = p$ and $p_{2i} = i/n$.

Example. We saw previously that admissible designs in polynomial regression models have the property of maximizing the $2pth$ moment subject to given values of the preceding moments. Such "upper principal representations" have a clean property with regard to their $2pth$ canonical moment: the $2pth$ canonical moment equals 1.

There is a wealth of information on canonical moments and connections to optimal designs in Lau and Studden (1985).

13.2. Orthogonal polynomials.

13.2.1. Relation to optimal design. The close relationship of various systems of orthogonal polynomials to optimal designs was seen in section 6.2. Generally, the points in the support of classical and Bayes optimal designs according to some alphabetic criteria coincide with the points of peak of various orthogonal polynomials. Thus it was seen that for extrapolation problems in polynomial regression, the c-optimal design is always supported at the peaks of the pth Chebyshev polynomial $T_p(x)$, the D-optimal design is supported at the turning points of Legendre polynomials plus the endpoints, and E-optimal designs concentrate on the turning points of Chebyshev polynomials as well. We will later give a list of the first few standard systems of orthogonal polynomials, and also a general algorithm for producing the entire sequence, which can be used on a computer to evaluate the turning points for any particular order p; note that p in this context coincides with the

degree of the polynomial regression model.

13.2.2. Basic properties. Orthogonal polynomials arise out of the following familiar approximation problem: there is a continuous function f defined on an interval $[a, b]$, and there is a finite dimensional subspace $\mathcal{A}$ of $C[a, b]$, and one wants to find an element p^* in $\mathcal{A}$ giving the smallest value of $\int (f(x) - p(x))^2 \, w(x)dx$, where $w(.)$ is a fixed weight function on $[a, b]$. By a weight function, one usually means a nonnegative integrable function. This is the usual leastsquares problem.

It is natural to write the solution p^* using the elements of a basis for $\mathcal{A}$; orthogonal polynomials essentially correspond to an orthogonal basis for $\mathcal{A}$. Thus, if $\mathcal{A}$ is of dimension $n+1$, then a sequence $\{\phi_i, 0 \leq i \leq n\}$ is an orthogonal basis if $\{\phi_i\}$ are linearly independent, and satisfy the inner product condition $\int \phi_i(x)\phi_j(x)w(x)dx = 0$ whenever $i \neq j$. Using elementary linear algebra, one can see that then the solution p^* has the representation

$$p^*(x) = \sum_{i=0}^{n} c_i \phi_i(x),$$

where $c_i = (\phi_i, f)/\|\phi_i\|^2$

where $(,)$ denotes inner product in the $L^2(w)$ space and $\|g\|^2 = (g, g)$.

There is another way to look at this; regardless of the least squares problem, one can form the expansion

$$\hat{f}_n(x) = \sum_{i=0}^{n} c_i f_i(x),$$

with c_i as above. One would intuitively expect that as subspaces of larger dimension are used, i.e., as n increases, the function $\hat{f}_n$ should approximate f more closely. There is a rich and long history of this method, variously known as orthogonal or Fourier expansions. Although Fourier expansions need not in general converge or converge to the parent function everywhere even if they do themselves converge, fairly general L_2 approximation results are indeed valid for the Fourier expansions.

Theorem. Under the assumption of completeness, the finite Fourier expansion $\hat{f}_n$ of f converges in L_2 to f; in fact,

$$\|\hat{f}_n - f\|^2 = \sum_{i=n+1}^{\infty} c_i^2,$$

and $\sum_{i=0}^{\infty} c_i^2 < \infty$, so that $\sum_{i=n+1}^{\infty} c_i^2 \to 0$.

13.2.3. Recursions and roots of orthogonal polynomials. The points at which certain orthogonal polynomials have zero derivatives are often the points in support of optimal designs. Therefore, it is important to know how orthogonal polynomials are found. A straightforward method would be to take an arbitrary basis and use the familiar Gram-Schmidt orthogonalization process. This, however, is unnecessary due to a remarkable recursion relation orthogonal polynomials satisfy.

Theorem. Define the first orthogonal polynomial as $\phi_0(x) = 1$;
define $\alpha_0 = \int xw(x)dx \,/\, \int w(x)dx$.

Define $\phi_1(x) = x - \alpha_0$. For $j > 1$, define recursively

$$\alpha_j = (\phi_j, x\phi_j)/\|\phi_j\|^2,$$
$$\beta_j = \|\phi_j\|^2/\|\phi_{j-1}\|^2,$$

and $\phi_{j+1}(x) = (x - \alpha_j)\,\phi_j(x) - \beta_j\phi_{j-1}(x)$.

Then $\{\phi_i\}$ is a sequence of orthogonal polynomials with respect to the inner product $(f,g) = \int f(x)g(x)w(x)dx$.

This three term recursion considerably simplifies the calculation of orthogonal polynomials for large values of n. Of course, in practice, relatively small n and standard weight functions $w(x)$ may be used, in which case the corresponding orthogonal polynomials are widely available. We will see such examples in the next subsection.

We close with two facts about the roots of orthogonal polynomials.

Theorem. Suppose $\{\phi_i\}$ is a system of orthogonal polynomials in an inner product space $L^2(w,[a,b])$. Then ϕ_k has exactly k roots which are real, simple, and in the interior of [a,b]. Furthermore, between two successive roots of ϕ_{k-1}, there is one root of ϕ_k.

13.2.4. Special orthogonal polynomials. The system of orthogonal polynomials $\{\phi_k\}$ are determined by the weight function $w(x)$; a special important case in applications is when the $(n+1)$ dimensional subspace in the general theory is the set of algebraic polynomials of degree n. In this important case, the orthogonal polynomials for certain standard weight functions are explicitly known and have been studied in great depth for their properties.

(a.) $a = -1,\ b = 1,\ w(x) = 1/\sqrt{(1-x^2)}$;

in this case, the orthogonal polynomials are Chebyshev polynomials $\{T_n(x)\}$; $T_n\ (x)$ also has the trigonometric interpretation that $T_n\ (cos\theta) = cos(n\theta)$.

The coefficients of the various powers of x in any $T_n(x)$ are explicitly known; one can see chapter 1 in Rivlin (1990). The first few Chebyshev polynomials are as follows:

n	$T_n\ (x)$
0	1
1	x
2	$2x^2 - 1$
3	$4x^3 - 3x$
4	$8x^4 - 8x^2 + 1$
5	$16x^5 - 20x^3 + 5$

(b.) $a = -1,\ b = 1,\ w(x) = 1$;

in this case, the orthogonal polynomials are the Legendre polynomials $\{P_n(x)\}$. Again, the exact coefficients of the various powers of x are known: one can see Rivlin (1969). The first few Legendre polynomials are as follows:

n	$P_n\ (x)$
0	1
1	x
2	$3x^2 - 1$
3	$5x^3 - 3x$
4	$35x^4 - 30x^2 + 3$
5	$63x^5 - 70x^3 + 15x$

(c.) $a = 0,\ b = \infty,\ w(x) = e^{-x}$;

in this case, the orthogonal polynomials are the Laguerre polynomials $\{L_n(x)\}$. One can see Gradshteyn and Ryzhik (1980) for the exact coefficients for the general degree. The first few Laguerre polynomials are as follows:

n	$L_n(x)$
0	1
1	x
2	$x^2 - 4x + 2$
3	$x^3 - 9x^2 + 18x - 6$
4	$x^4 - 16x^3 + 72x^2 - 96x + 24$
5	$x^5 - 25x^4 + 200x^3 - 600x^2 + 600x - 120$

(d.) $a = -\infty$, $b = \infty$, $w(x) = e^{-x^2}$;

in this case, the orthogonal polynomials are the very familiar Hermite polynomials. The exact coefficients are again available in Gradshteyn and Ryzhik (1980), and the first five Hermite polynomials are the following:

n	$H_n(x)$
0	1
1	x
2	$2x^2 - 1$
3	$2x^3 - 3x$
4	$4x^4 - 12x^2 + 3$
5	$4x^5 - 20x^3 + 15x$

Other important cases include the symmetric Beta function $w(x) = x^{\alpha-1}(1-x)^{\alpha-1}$, for $\alpha > 0$, in which case one gets the Ultraspherical polynomials, and the general Beta function $w(x) = x^{\alpha-1}(1-x)^{\beta-1}$, for $\alpha, \beta > 0$, and one gets the Jacobi polynomials.

There is a unifying feature regarding all of these cases: it is that there is a function $u(x)$ such that the nth orthogonal polynomial $\phi_n(x)$ admits the representation

$$\phi_n = u^{(n)}(x)/w(x),$$

where $u^{(n)}$ denotes the nth derivative of u. This is sometimes called Rodriguez's formula and the correct choice of $u(x)$ is known for each case.

13.2.5. Linearization of nonlinear functions. A major use of orthogonal polynomials is in approximating complicated functions by linear combinations of orthogonal polynomials. Note that the L^2 theory only assures close approximation in $L2$ norm, but in optimal design, one may need an assurance of uniform approximation in the design space. Some very interesting results are indeed known, and we think they are useful in linearization of nonlinear models. We mention two of these results.

Theorem. (Dini-Lipschitz). Suppose f is a continuous function on a bounded interval $[a, b]$ and let $w(f, .)$ denote its modulus of continuity. Suppose $s_n(f)$ is the nth partial sum in the Chebyshev expansion of f; if $w(f, 1/n)\ \log\ n \to 0$ as $n \to\ \infty$, then $s_n(f)\ \to f$ uniformly.

Corollary. If f is Lipschitz (α) for $\alpha > 0$, then the Chebyshev expansion of f uniformly converges to f.

In addition to the above theorem, for purposes of deciding how many terms one should use, estimates of the error are useful. There are several results known; we find the following useful.

Theorem. Let $E_n(f)$ denote the error in the approximation of f by $s_n(f)$ using supnorm, and let $E_n^*(f)$ denote the same error by using the best polynomial approximation to f in supnorm. Then

$$E_n(f)\ \leq\ 4(1 + \frac{log\ n}{\pi^2})\ E_n^*(f).$$

The suggestion in Atkinson and Donev (1992) is to linearize a nonlinear model by using its Taylor expansion; we believe that Chebyshev expansions can estimate more efficiently with a smaller number of terms. One reason is the following theorem:

Theorem. Consider the expansion of a continuous function in terms of ultraspherical polynomials defined in section 13.3.3. Then, the choice $\alpha = 1/2$ always gives the best

approximation in *sup* norm if the coefficients $\{a_i, i > n\}$ in the ultraspherical expansion of f are nonnegative for that given n; in particular, a Chebyshev expansion corresponding to $\alpha = \frac{1}{2}$ is better than a Taylor expansion which corresponds to $\alpha = \infty$.

Acknowledgement

I learned from Bill Studden the little I know about optimal designs. I am very thankful for the scholarly inspiration he provides, in an understated brilliant way.

References

Akhiezer, N. (1962). Some questions in the theory of moments, AMS, Providence, Rhode Island.

Antelman, G.R. (1965). *Insensitivity to non-optimal design in Bayesian decision theory*, J. Amer. Statist. Assoc., **60**, pp. 584-601.

Atkinson, A.C. and Donev, A.N. (1992). *Optimum experimental designs.* Clarendon Press, Oxford.

Basu, S. and DasGupta, A. (1992). Robustness of standard confidence intervals under departure from normality, To appear in Ann. Stat.

Berg, C. (1985). *On the preservation of determinacy under convolution*, Proc. Amer. Math. Soc. **93**, 351-357.

Berg, C. (1995). Indeterminate moment problems and the theory of entire functions, *Jour. Comp. Appl. Math.*, **65**, 27-55.

Berger, J. (1986). *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. Springer, New York.

Berger, J. (1994). An overview of Robust Bayesian analysis, Test, 3, No. 1, 5-59.

Bickel, P.J., and Herzberg, A.M. (1979). Robustness of design against autocorrelation in time I, *Ann. Statist.*, **7**, pp. 77-95.

Billingsley, P. (1986). Probability and measure, John Wiley, New York.

Bock, J. and Toutenberg, H. (1991). Sample size determination in clinical research, in Handbook of Statistics, **vol. 8**, 515-538.

Bose, R.C. (1948). "The design of experiments". In *Proceedings of the Thirty-Fourth Indian Science Congress, Delhi 1947.* Indian Science Congress Association, Calcutta, (1)-(25).

Bowman, K.O. and Kastenbaum, M.A. (1975). Sample size requirement: single and double classification experiments, in selected tables in mathematical statistics (vol. 3), edited by IMS, AMS, Providence, Rhode Island.

Box, G.E.P. and Draper, N.R. (1987). *Empirical model-building and response surfaces.* Wiley, New York.

Box, G.E.P. and Hunter, J.S. (1957). "Multi-factor experimental designs for exploring response surfaces." *Annals of Mathematical Statistics* 28, 195-241.

Box, G.E.P. and Lucas, H.L. (1959). Design of experiments in non-linear situations. *Biometrika*, **46**, 77-90.

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**, 217-221.

Brown, L.D. (1986). Fudnamentals of statistical exponential families, IMS Lecture notes - monograph series, **vol - 9**, Hayward, California.

Brown, L.D. (1991). Minimaxity, more or less, In Stat. Dec. Theory and Related Topics, V, Eds. S. Gupta and J. Berger, 1-18.

Caselton, W.F. and Zidek, J.V. (1984). Optimal monitoring network designs. *Statist. Probab. Lett.* **2** 223-227.

Chaloner, K. (1984). "Optimal Bayesian experimental design for linear models." *Annals*

of Statistics 12, 283-300. "Correction". *Ibidem 13*, 836.

Chaloner, K. (1989). Bayesian design for estimating the turning point of a quadratic regression. *Communications in Statistics, Theory and Methods,* **18(4)**, 1385-1400.

Chaloner, K. (1993). A note on optimal Bayesian design for nonlinear problems. *Jour. Statist. Planning and Inference*, **37**, 229-235.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian Designs applied to Logistic Regression Experiments. *Jour. Statist. Planning Inference*, **21**, 191-208.

Chaloner, K. and Larntz, K. (1986). Optimal Bayesian Designs applied to Logistic Regression Experiments. *University of Minnesota Technical Report.*

Chaloner, K. and Verdinelli, I. (1994). Bayesian experimental design: a review, Technical Report, Department of Statistics, University of Minnesota.

Cheng, C.-S. (1978b). "Optimal designs for the elimination of multi-way heterogeneity." *Annals of Statistics 6*, 1262-1272.

Chernoff, H. (1953). "Locally optimum designs for estimating parameters", *Ann. Math. Statist.*, **24**, 586-602.

Chernoff, H. (1972). *Sequential analysis and optimal design.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

Clyde, M.A. (1993). An object-oriented system for Bayesian nonlinear design using xlisp-stat. University of Minnesota, School of Statistics, Technical Report 587.

DasGupta, A., Mukhopadhyay, S. and Studden, W.J. (1992). Compromise designs in heteroscedastic linear models. *Jour. Statist. Planning and Inference*, **32**, 363-384.

DasGupta, A. and Mukhopadhyay, S. (1988). Uniform and subuniform posterior robustness: sample size problem, Tech. Report, Purdue University.

DasGupta, A. and Studden, W.J. (1988). Robust Bayesian analysis and optimal experimental designs in normal linear models with many parameters. I. Technical Report, Dept. of Statistics, Purdue University.

DasGupta, A. and Studden, W.J. (1991). Robust Bayes Designs in Normal Linear Models.

Ann. Statist., **19**, 1244-1256.

DasGupta, A. and Mukhopadhyay, S. (1994). Uniform and subunivorm posterior robustness: the sample size problem, Proc. of the 1st International workshop on Bayesian robustness, Special issue of *Jour. Stat. Planning and Inf*, **40**, 189-204.

DasGupta, A. and Vidakovic, B. (1994). Sample sizes in ANOVA: The Bayesian point of view, Tech. Report, Purdue University, submitted *Jour. Stat. Planning and Inf.*

DasGupta, A. and Zen, M. M. (1996). Bayesian bioassay design, Tech. Report, Purdue University, Submitted *Jour. Stat. Planning and Inf.*

Dehnad, K. (ed.) (1989). *Quality control, robust design, and the Taguchi method.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

DeRobertis, L. and Hartigan, J.A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235-244.

Dette, H. (1991). A note on robust designs for polynomial regression. *Jour. Statist. Planning and Inference,* **28**, 223-232.

Dette, H. (1992). Optimal designs for a class of polynomials of odd or even degree, Ann. Statist., **20**, 238-259.

Dette, H. (1993a). Elfving's theorem for D-optimality, *The Annals of Statist.*, **21**, 753-766.

Dette, H. (1993b). A note on Bayesian c- and D-optimal designs in nonlinear regression models. *Manuscript.*

Dette, H. and Neugebauer, H.-M. (1993). Bayesian D-optimal designs for exponential regression models. To appear in *Jour. Stat. Planning and Inference.*

Dette, H. and Sperlich, S. (1994a). Some applications of continued fractions in the construction of optimal designs for nonlinear regression models. *Manuscript.*

Dette, H. and Sperlich, S. (1994b). A note on Bayesian D-optimal designs forr general exponential growth models. *Manuscript.*

Dette, H. and Studden, W.J. (1994a). Optimal designs for polynomial regression when the degree is not known, Technical Report, Purdue University.

Dette, H. and Studden, W.J. (1994b). A geometric solution of the Bayes E-optimal design problem, Stat. Dec. Theory and Related Topics V, Eds. S. Gupta and J. Berger, 157-170.

Diaconis, P. (1987). Bayesian numerical analysis, In *Stat. Dec. Theory and Related Topics*, Eds. S. Gupta and J. Berger, IV, **Vol. 1**, 163-176.

Diaconis, P. (1987). Application of the method of moments in probability and statistics, in Moments in Mathematics, AMS, Providence, Rhode Island.

Donev, A.N. (1988). *The construction of exact D-optimum experimental designs.* Ph.D. Thesis, University of London.

Dykstra, Otto, Jr. (1971). The augmentation of experimental data to maximize $|X^TX|$. *Technometrics* **13** 682-688.

Eaton, M.L., Giovagnoli, A. and Sebastiani, P. (1994). A predictive approach to the Bayesian design problem with application to normal regression models. *Technical Report 598*, School of Statistics, University of Minnesota.

Elfving, G. (1952). "Optimum allocation in linear regression theory." *Annals of Mathematical Statistics 23*, 255-262.

El-Krunz, S.M. and Studden, W.J. (1991). "Bayesian optimal designs for linear regression models." *Annals of Statistics 19*, 2183-2208.

Elliott, P.D.T.A. (1979). Probabilistic Number Theory, **Volume II**, Springer-Verlag, New York.

Erdos, P. (1958). Problems and results on the theory of interpolation, I, *Acta. Math. Acad. Sci. Hungar.*, **9**, 381-388.

Farrell, R.H., Kiefer, J. and Walbran, A. (1967). Optimum multivariate designs. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA 1965 and 1966,* **Volume 1** (Eds. L.M. Le Cam, J. Neyman). University of California, Berkeley, CA.

Fedorov, V.V. (1972). *Theory of optimal experiments.* Academic Press, New York.

Ferguson, T.S. (1989). Who solved the secretary problem?, Statistical Science, 4, No. 3, 282-296.

Fisher, R.A. (1949). Design of Experiments, Hafner, New York.

Freeman, P.R. (1983). The secretary problem and its extensions - a review. Internat. Statist. Rev., 51, 189-206.

Friedman, M. & Savage, L.J. (1947). "Experimental determination of the maximum of a function", *Selected Techniques of Statistical Analysis,* 363-372. New York, McGraw-Hill.

Gaffke, N. and Krafft, O. (1982). "Exact D-optimum designs for quadratic regression." *Journal of the Royal Statistical Society Series B*, 44, 394-397.

Ghosh, S. (ed.)(1990). *Statistical design and analysis of industrial experiments.* Marcel Dekker, New York.

Ghosh, M., Sinha, Bimal K. and Mukhopadhyay, N. (1976). Multivariate sequential point estimation, Jour. Mult. Analysis, **6**, 281-294.

Giovagnoli, A. and Verdinelli, I. (1985). Optimal block designs under a Hierarchical linear model, in *Bayesian Statistics 2.* J.M. Bernardo et al. eds. North Holland.

Giovagnoli, A. and Verdinelli, I. (1983). Bayes D-optimal and E-optimal block designs. *Biometrika*, **70**, 3, 695-706.

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

Gradshteyn, I.S. and Ryzhik, I.M. (1980). Table of Integrals, Series and Products, Academic Press, New York.

Haines, L.M. (1987). The application of the annealing algorithm to the construction of exact D-optimum designs for linear-regression models. *Technometrics* **29**, 439-47.

Hedayat, A.S., Jacroux, M. and Majumdar, D. (1988). "Optimal designs for comparing test treatments with a control." *Statistical Science 3*, 462-476. "Discussion." *Ibidem*, 477-491.

Herzberg, A.M. and Cox, D.R. (1969). "Recent work on the design of experiments: A bibiliography and a review." *Journal of the Royal Statistical Society Series A* **132**, 29-67.

Huber, P.J. (1972). Robust statistics: A review, *Ann. Math. Statist.*, **43**, pp. 1041-1067.

Huber, P.J. (1975). Robustness and designs, in: *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, Ed., North Holland, Amsterdam.

Huber, P.J. (1981). Robust statistics, Wiley, New York.

Joseph, L., Wolfson, D. and Berger, R. (1994). Some comments on Bayesian sample size determination. Preprint.

Joseph, L. and Berger, R. (1994). Bayesian sample size methodology with an illustration to the difference between two binomial proportions. Preprint.

Kacker, R.N. (1985). Off-line quality control, parameter design, and the Taguchi method, Jour. Qual. Tech., 17, 176-188.

Karlin, S. & Shapley, L.S. (1953). *Geometry of Moment Spaces,* **vol. 12**, of Amer. Math. Soc. Memoirs.

Karlin, S. and Studden, W.J. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics.* Interscience, New York.

Kemperman, J.H.B. (1968). *The general moment problem, a geometric approach*, Annals Math. Statist. **39**, 93-122.

Kemperman, J.H.B. (1972). *On a class of moment problems*, Proc. Sixth Berkeley Symposium on Math. Statist. and Prob. **2**, 101-126.

Kiefer, J. (1953). "Sequential minimax search for a maximum", *Proc. Amer. Math. Soc.*, **4**, 502-506.

Kiefer, J.C. (1959). "Optimum experimental designs". *Journal of the Royal Statistical Society Series B*, **21**, 272-304. "Discussion on Dr. Kiefer's paper". *Ibidem,* 304-319.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs on regression problems. *Ann. Math. Statist.* **30**, 271-94.

Kiefer, J.C. (1974). "General equivalence theory for optimum designs (approximate theory)." *Annals of Statistics*, **2**, 849-879.

Kiefer, J.C. and Studden, W.J. (1976). "Optimal designs for large degree polynomial regression." *Annals of Statistics*, **4**, 1113-1123.

Kiefer, J. (1987). Introduction to Statistical Inference, Springer-Verlag, New York.

Korner, T.W. (1989). Fourier analysis, Cambridge, New York.

Krein, M.G. and Nudelman, A.A. (1977). The Markov moment problem and extremal problems, AMS, Providence, Rhode Island.

Kurotschka, V. (1978). "Optimal design of complex experiments with qualitative factors of influence." *Communications in Statistics, Theory and Methods A7*, 1363-1378.

Lau, T.-S, and Studden, W.J. (1985). Optimal designs for trigonometric and polynomial regression using canonical moments." *Annals of Statistics*, **13**, 383-394.

Leamer, E.E. (1978). *Specification Searches: Ad hoc Inference with Nonexperimental Data.* Wiley, New York.

Lee, C.M.-S (1988). "Constrained optimal designs." *Journal of Statistical Planning and Inference*, **18**, 377-389.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York.

Lindley, D.V. (1956). "On a measure of the information provided by an experiment." *Annals of Mathematical Statistics*, **27**, 986-1005.

Majumdar, D. (1992). Optimal designs for comparing test treatments with a control using prior information, Ann. Statist., **20**, 216-237.

Majumdar, D. (1995). Optimal and efficient treatment-control designs, Preprint.

Marcus, M.B. and Sacks, J. (1976). Robust design for regression problems, In Stat. Dec. Theory and Related Topics, II, Eds. S. Gupta and D. Moore, 245-268.

Mitchell, T., Sacks, J. and Ylvisaker, D. (1994) Asymptotic Bayes criteria for nonparametric response surface design, *Ann. Statist.*, **22**, 634-651.

Mukhopadhyay, S. and Haines, L. (1993). Bayesian D-optimal designs for the exponential growth model. To appear in *Jour. Stat. Planning and Inference.*

Nalimov, V.V. (1974). "Systematization and codification of the experimental designs–The survey of the works of Soviet statisticians." In *Progress in Statistics. European Meeting of Statisticians, Budapest 1972,* **Volume 2** (Eds. J. Gani, K. Sarkadi, I. Vincze). Colloquia Mathematica Societatis János Bolyai 9, North-Holland, Amsterdam, 565-581.

Nalimov, V.V. (ed.)(1982). *Tables for planning experiments for factorials and polynomial models.* (in Russian), Metallurgica, Moscow.

Odeh, R.E. (1975). Sample size choice: charts for experiments with linear models, Marcel Dekker, New York.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *Jour. Roy. Statist. Soc., Ser. B,* **40**, 1-41.

Owen, R.J. (1970). The optimum design of a two-factor experiment using prior information *Ann. Math. Statist.* **41**, 1917-34.

Papalambros, P.Y. and Wilde, D.J. (1988). Principles of optimal design, Cambridge, New York.

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

Pilz, J. (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models.* Wiley, New York.

Polasek, W. (1985). Sensitivity analysis for general and hierarchical linear regression models. In *Bayesian Inference and Decision Techniques with Applications* (P.K. Goel and A. Zellner, eds.) 375-387. North-Holland, Amsterdam.

Powell, M.J.D. (1981). Approximation theory and methods, Cambridge University Press, New York.

Pukelsheim, F. (1980). "On linear regression designs which maximize information." *Journal of Statistical Planning and Inference 4*, 339-364.

Pukelsheim, F. (1988). Analysis of variability by analysis of variance, in optimal design and analysis of experiments, eds. Y. Dodge, V.V. Fedorov and H.P. Wynn, North-Holland, New York.

Pukelsheim, F. and Rieder, S. (1992). "Efficient rounding of approximate designs." *Biometrika 79*, 763-770.

Pukelsheim, F. (1993). *Optimal Design of Experiments.* Wiley, New York.

Pukelsheim, F. and Studden, W.J. (1993). "E-optimal designs for polynomial regression." *Annals of Statistics 21*, No. 1.

Rao, C.R. (1946). Difference sets and combinatorial arrangements derivable from finite geometries. *Proc. Nat. Inst. Sc.* **12**, 123-135.

Rao, C.R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *J. Roy. Statist. Soc.* **B9**, 128-140.

Rao, C.R. (1973). *Linear statistical inference and its applications* (2nd edn). Wiley, New York.

Rivlin, T.J. (1969). An introduction to the approximation of functions, Dover, New York.

Rivlin, T.J. (1990). Chebyshev polynomials, Second edition, Wiley-Interscience, New York.

Royden, H.L. (1953). *Bounds on a distribution function when its first n moments are given*, Annals Math. Statist. **24**, 361-376.

Rubin, H. (1977). Robust Bayesian estimation, In Stat. Dec. Theory and Related Topics, II, Eds. S. Gupta and D. Moore, 351-356.

Sacks, J., and Schiller, S. (1988). Spatial designs, in *Statistical Decision Theory and Related Topics*, **Vol. IV**. Springer-Verlag, New York, pp.385-399.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4**, 409-35.

Sacks, J. and Ylvisaker, D. (1970). Statistical designs and integral approximation. *Proc. of the 12th Bien. Seminar Canad. Math. Cong.*, 115-136. Montreal.

Sacks, J., and Ylvisaker, D. (1964). Designs for Regression Problems with correlated errors III, *Ann. Math. Statist.*, **41**, 2057-2074.

Sarkadi, K. and Vincze, I. (1974). Mathematical methods of statistical quality control, Academic Press, New York.

Schoenberg, I. J. (1959). On the maximization of some Hankel determiants and zeros of classical orthogonal polynomials, *Indag. Mathematical,* **21**, 282-290.

Schumacher, P. and Zidek, J.V. (1993). Using prior information in designing intervention detection experiments, Ann. Stat., **21**, 447-463.

Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions, *Ann. Math. Statist.,* **33**, 224-236.

Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear regression.* Wiley, New York.

Shah, K.S. and Sinha, B.K. (1989). *Theory of optimal designs.* Lecture notes in statstics 54, Springer, New York.

Siegmund, D. (1985). Sequential analysis, Springer-Verlag, New York.

Silvey, S.D. (1980). *Optimal design.* Chapman and Hall, London.

Skibinsky, M. (1968). Extreme nth moments for distributions on [0,1] and the inverse of a moment space map. *J. Appl. Probab.* **5**, 693-701.

Smith, A.F.M. and Verdinelli, I. (1980). A note on Bayesian design for inference using a Hierarchical linear model. *Biometrika*, **67**, 613-619.

Smith, K. (1918). "On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations", *Biometrika*, **12**, 1-85.

Spiegelhalter, D.J. and Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medecine*, **5**, 1-13.

Starr, N. and Woodroofe, M.B. (1969) *Remarks on sequential point estimation*, Proc. Nat. Acad. Sci., **63**, pp. 285-288.

Staudte, R.G. and Sheather, S.J. (1990). Robust estimation and testing, Wiley-

Interscience, New York.

Steinberg, D.M. and Hunter, W.G., (1984). Experimental design: review and comment. *Technometrics*, **26**, 71-97.

Stigler, S.M.(1971). "Optimal experimental design for polynomial regression." *Journal of the American Statistical Association 66,* 311-318.

Stigler, S.M. (1974). Gergonne's 1815 paper on the design and analysis of polynomial regression experiments, Historia Mathematica, 1, 431-447.

Studden, W.J. (1977). "Optimal designs for integrated variance in polynomial regression." In *Statistical Decision Theory and Related Topics II. Proceedings of a Symposium, Purdue University 1976* (Eds. S.S. Gupta, D.S. Moore). Academic Press, New York, 411-420.

Studden, W.J. (1982). "Some robust-type D-optimal designs in polynomial regression." *Journal of the American Statistical Association 77*, 916-921.

Tang, Dei-in (1993). Minimax regression designs under uniform departure models, Ann. Stat., **21**, 434-446.

Toman, B. and Notz, W. (1991). Bayesian optimal experimental design for treatment control comparisons in the presence of two-way heterogeneity. *J. Statist. Plann. Infer.*, **27**, 51-63.

Toman, B. (1992). Bayesian robust experimental designs for the one-way anaylsis of variance. *Statistics and Probability Letters*, **15**, 395-400.

Toman, B. and Gastwirth, J.L. (1993). Robust Bayesian experimental design and estimation for analysis of variance models using a class of normal mixtures. *Journal of Statistical Planning and Inference*, **35**, 383-398.

Verdinelli, I. (1983). Computing Bayes D- and E-optimal designs for a two-way model. *The Statistician*, **32**, 161-167.

Verdinelli, I. and Wynn, H.P. (1988). Target attainment and experimental design, a Bayesian approach, In optimal Design and analysis of experiments, Eds. Y. Dodge, V.v. Dedorov and H.P. Wynn, North-Holland, New York.

Wald, A. (1943). "On the efficient design of statistical investigations." *Annals of Mathematical Statistics 14*, 134-140.

Wasserman, L. (1992). Recent methodological advances in Robust Bayesian inference, In Bayesian Statistics, 4, Eds. J. Bernardo et al, Oxford University Press.

Whittle, P. (1973). "Some general points in the theory of optimal experimental design." *Journal of the Royal Statistical Society Series B 35*, 123-130.

Wilde, D.J. (1978). Globally optimal design, Wiley - Interscience, New York.

Wu, C-F. (1988). Optimal design for percentile estimation of a quantal response curve. In *Optimal Design and Analysis of Experiments*, eds. Y. Dodge, V.V. Fedorov and H.P. Wynn, North-Holland, Amsterdam.

Wynn, H.P. (1977). "Optimum designs for finite populations sampling." In *Statistical Decision Theory and Related Topics II. Proceedings of a Symposium, Purdue University 1976* (Eds. S.S. Gupta, D.S. Moore). Academic Press, New York, 471-478.

Wynn, H.P. (1984). Jack Kiefer's contributions to experimental design. *Ann. Statist.* **12**, 416-23.