BAYESIAN DESIGNS FROM ASYMPTOTICS

DON YLVISAKER

Statistics, UCLA

ABSTRACT. The paper sets forth a Bayesian nonparametric regression framework in which the principal issue is one of where observations should be taken. Specific criteria are advanced to measure information gained so that designs might be compared, and asymptotics are introduced so that problems found in this way are more manageable. Illustrations are drawn from different settings, and some remarks are made about the general program

0. Introduction.

This is a paper in the optimal design tradition, begun with such prominent works as Kiefer ((1958) and (1959)) and Kiefer and Wolfowitz ((1959) and (1960)). Beyond review articles detailing developments in the area, Atkinson and Fedorov (1989), Bandemer, Nather and Pilz (1987), and Steinberg and Hunter (1984) for example, there are the books by Fedorov (1972), Pilz (1991) and Pukelsheim (1993), amongst others. Much of the research has featured observation of a parametric response surface in the presence of independent errors, posing the question of where should one observe to gain maximum information about the parameters of the model. Some of this work has reached a high level of mathematical sophistication, exemplified by Cheng (1978) in the case of discrete problems, and by Dette and Studden (1997) in continuous ones.

At the same time, the inherent difficulty of the accompanying optimization problems has led to work meant, in part, to ease the amount of labor involved. One method for doing this is to adopt a Bayesian stance - for practical, if not philosophical, reasons. See O'Hagan (1978), Sacks and Ylvisaker (1985) and Ylvisaker (1987) about this. A concomitant benefit of this approach is that one can downplay parametric models in the process, allowing the response surface to take a freer form. A second device used to soften difficulty is asymptotics, see Mitchell, Sacks and Ylvisaker (1994) in this regard especially, henceforth MSY, and note that asymptotics is given a broader meaning there than that conveyed by the notion of large sample sizes.

Even with the reduction in difficulty that comes from Bayesian asymptotics, resulting problems are nontrivial. To bring in specific items of interest, we begin with the setting employed in MSY. In words, the asymptotics stem from imagining a weakly varying signal (the response surface) in the presence of noise (the observational error); the Bayesian nonparametric slant comes from viewing the response surface as a sample outcome of a stochastic process. The set of sites for observation can be taken in some generality and the question of which sites to observe, and how often, is posed in light of Bayesian criteria for information gain. The problems that come about in this way have a strong game theory flavor to them.

The necessary framework is set out in Section 1, where the results of MSY are given some review and broadened a little. Thus, given the ability to observe a stochastic process with error, problems of finding designs termed A, D or G-optimum are posed. Some results are given in Section 2. There we go through a sequence of settings in which general matters can be made more precise and some explicit designs can be provided for the sake of intuition. In particular, the question of D-optimum designs is related to the older problems of estimating the unknown mean of a stochastic process via a linear estimator, see especially the work of Hajek ((1956) and (1962)) and Ylvisaker (1964) in this direction. Lastly some remarks are made about the worthiness of designs, produced through asymptotics, in the finite domain.

Tom Ferguson and I have been colleagues at UCLA for more than thirty years, advantage to me. He is the epitome of a gentleman scholar, and is hugely influential in his teaching of others. It is a pleasure then to at least name some of his wide-ranging interests in this paper: Bayes procedures, nonparametrics, asymptotics, game theory. In the same vein, apologies are offered over the mention of a minimum variance unbiased estimator - it seemed to be a part of the story.

1. The Design Setup.

The set T will consist of sites at which one might draw observations and, in our concrete discussion, T might be 2^k or an interval in R^d, for example. An observation at t in T has the form

$$Y(t) = \beta + X(t) + \varepsilon, \tag{1}$$

where β is a normal variable with mean 0 and variance v, X is a Gaussian process indexed by T and having mean 0 and covariance function R, while the ε error terms are independent normals with mean 0 and variance σ^2. In this framework we think of our uncertainty about the response surface X as being modelled by the process description given.

Despite the level of generality indicated one can state a basic question: where should one take an alloted n observations, replication being allowed? The question is made precise once a criterion is stated whereby we can decide between proposed answers. To maintain simplicity for this let T be a finite set - it is normally so in real practice and approximately so in any event. Here are three criteria for comparing different designs: look to the process X conditioned on the observations drawn, and state a preference for small values of

(2)
D : the generalized variance of the conditioned process;
G : the maximum variance over T of the conditioned process;
A : the average variance over T of the conditioned process.

(The last of these is well-specified once we give some meaning to "average".)

These problems are simply stated, but cannot be solved in any kind of generality. Invoking the following asymptotics brings them to a more manageable form. Think of a "weakly varying signal" (here X) in the presence of "noise" (observation error). More precisely, take R to be fixed and allow v and σ^2 to tend to infinity in ratio $\gamma = v/\sigma^2$. One can then translate the criteria listed above to more concrete forms by following (3.2) to (3.9) of MSY. Let ξ be the design measure that puts weight $1/n$ at each chosen t, replications being allowed. Minimize, by choice of ξ, the criteria

$$
(3)\qquad
\begin{aligned}
D:&\quad \lambda \iint R(s,t)d\xi(s)d\xi(t) - \int R(t,t)d\xi(t),\\
G:&\quad \max_u \left\{\lambda \iint R(s,t)d\xi(s)d\xi(t) - 2\int R(u,t)d\xi(t)\right\},\\
A:&\quad \lambda \iint R(s,t)d\xi(s)d\xi(t) - 2\iint R(s,t)d\pi(s)d\xi(t),
\end{aligned}
$$

where $\lambda = (n\gamma/(1+n\gamma))$ and, for A-optimality, π is some probability measure on T.

A standard device in the literature, going back to Kiefer and Wolfowitz, allows the design measure ξ to be general rather than insist that $n\xi$ be integral. One speaks then of the *approximate theory* to constrast it with an *exact theory* that expects to produce directly implementable designs. We mostly follow the approximate theory here.

Some transparency can be given to the problems in (3) if one takes X to have constant variance over T, takes γ to be 0, or both. We set forth simplified versions of each criterion in (3) as

$$\begin{aligned} &D: \quad \min_{\xi} \iint R(s,t)d\xi(s)d\xi(t), \\ (4) \qquad &G: \quad \max_{\xi}\min_{u} \int R(u,t)d\xi(t), \\ &A: \quad \max_{\xi} \iint R(s,t)d\pi(s)d\xi(t), \end{aligned}$$

where π is a fixed probability measure on T. In particular by now, the G-optimality problem is one of finding the optimum strategy for the maximizing player in a zero-sum game with payoff kernel R. The present A-optimality problem is that of finding the maximizing player's response in the knowledge of the minimizing player's strategy (indeed, A-optimum designs under (4) can sit on a single point). The D-optimality criterion seeks a common strategy for the two players in order to minimize the payout.

It is clear that D-optimum designs are more readily obtained than are G-optimum designs, a finding common to many settings. It is shown in MSY that the problem of minimizing the generalized variance of the conditioned process is the same as that of maximizing the generalized variance of the observations. Thus attention can focus on at most n, as opposed to card(T), variables. Dealing with A-optimality would normally be of intermediate difficulty, as in (3) if not in (4). We turn now to the task of expanding the results of MSY as they apply to problems found under (3) or the cleaner versions that are given at (4).

2. Optimum Designs.

The structure of the section is something like the following. We begin with D and A-optimality and with what the location of such designs entails. Then in four different settings a series of comments is made about the covariance functions looked at, some of the optimum designs that attach to them, and the relationships between optimum designs as they surface.

Consider D-optimality as at (3) and set $H(s) = \int R(s,t)d\xi(t)$. Then, since R is nonnegative definite, standard perturbation techniques show that ξ^* is optimum if and only if the function $2\lambda H^*(s) - R(s,s)$ achieves its minimum value at all s in the support of ξ^*. One arrives at the D-optimal condition in (4) by taking X to have constant variance over the set T, with $\lambda \neq 0$. In this case ξ^* is optimum if and only if the function H^* achieves its minimum value at all s in the support of ξ^*.

A-optimality at (3) can be viewed in the same way. Let $J(s) = \int R(s,t)d\pi(t)$. Here ξ^* is optimum if and only if $\lambda H^*(s) - J(s)$ achieves its minimum value at all s in the support of ξ^*. Of course at (4) we can see that any design is optimum provided it concentrates on the maximum of J.

Setting 1: Let $T = \{-1,1\}^k$, i.e., consider an experimental situation in which there are k factors each occurring at two levels, high and low say. The elements of T can be seen to form a transitive group under the operation of direct multiplication. With this in mind, a process X is called *stationary* if $R(m \cdot s,\ m \cdot t) = R(s,t)$ for all m, s and t in T.

Here are some correlation functions of stationary processes on T. Let ρ_i, $i = 1, \ldots, k$ be a collection of positive correlations, normalized so that $\Sigma_i\ \rho_i = 1$, say. Take

$$R(s,t) = \int \exp\left\{\theta \sum |s_i - t_i| \log\ \rho_i\right\} dF(\theta) \tag{5}$$

for any distribution function F on $\theta \geq 0$. If the ρ_i are constant, one has a sub-class of *isotropic* processes where the coordinates of X are exchangeable. (Correlations like this show up in Mitchell, Morris and Ylvisaker (1995), along with other information about the structure of stationary processes on T.)

1a) If one assumes X to be stationary, the uniform measure ξ^* on T is a D-optimum design. This remark requires nothing more than the observation that

$$H^*(s) = \sum_t R(s,t)2^{-k} = \sum_t R(m \cdot s,\ m \cdot t)2^{-k} = H^*(m \cdot s)$$

for all m in T. Hence H^* is constant over T, sufficient for D-optimality under (4). Then too, if π is uniform on T, $J(s)$ is constant over T and the uniform measure is A-optimum for this choice of weighting.

1b) Going further with this, the two person zero-sum game with symmetric payoff kernel R is in equilibrium if each player picks a site at random - the payoff is the constant value of H^*. Then the uniform design is G-optimum under the criteria at (3) as well. The sense of this comes from the uniform nature of the criterion, in opposition to A-optimality generally, but here is quick indication of why it is so. If ξ were G-better than the uniform measure ξ^*,

$$\max_u \left\{ \lambda \iint R(s,t)d\xi(s)d\xi(t) - 2\int R(u,t)d\xi(t) \right\}$$
$$= \lambda \iint R(s,t)d\xi(s)d\xi(t) - 2\min_u H(u) < (\lambda - 2)H^*.$$

But $\iint R(s,t)d\xi(s)d\xi(t) \geq H^*$ by the D-optimality of ξ^*, and $\min_u\ H(u) \leq H^*$ since ξ^* is a maximizing strategy for the game, yielding a contraction. (This argument has force whenever the D-optimum measure is supported on the whole of the design space T.)

1c) Of course we are begging the issue of implementation of the uniform design if the design calls for n observations and n is not a multiple of 2^k. For D-optimality in the exact theory one should be minimizing $\Sigma\Sigma\ R(t_i, t_j)$, the t_i being n possibly replicated sites in T.

Suppose, for instance, that n is less than 2^k. Asymptotics suggest an interesting, equivalent problem here. Take a correlation function of the form (5) and suppose that the distribution function does not put too much weight on small θ. Then the D-optimum design is a *maximin distance design.* This conclusion means that if the distance $d(s,t)$ between points s and t in T is taken to be $\{-\Sigma|s_i - t_i| \log \rho_i\}$, the D-optimum design has the property that the minimum distance between designs is maximized. See Johnson, Moore and Ylvisaker (1990) for this, as well as the following.

Take a correlation function of the form (5) and suppose that the distribution function does not put too much weight on small θ, then the G-optimum design is a *minimax distance design.* Thus if one checks the distance from the most remote site in T to the design, using the distance d above, such distance is minimized. (The second order term that arises in the asymptotics is also known for suitable expansions, and allows for some further refinements to these concepts.)

These alternative problems, while provocative in tone, require quite sophisticated computations. See Hardin and Sloane (1993) in this regard, for instance.

Setting 2: Let T be the interval $[-1,1]$ and take X to be a stationary process on T.

2a) Suppose $R(s,t) = \exp\{-\theta|s-t|\}$. Take $\xi^* = (\theta/(1+\theta)) \cup (-1,1) + (2(1+\theta))^{-1}(\partial_{-1}+\partial_1)$, where $\cup$ denotes the uniform measure and the ∂'s are point masses, and compute directly that $H^*(s) = \int R(s,t)d\xi^*(t)$ is identically $(1+\theta)^{-1}$. Accordingly a design that takes a proportion $(2(1+\theta))^{-1}$ of the n observations at each of the two endpoints, along with the remaining observations equally spaced on the interval, will be approximately D-optimum at (4). Since ξ^* has full support, the design is G-optimum as well. This measure was first given by Hajek (1956) in the context of unbiased estimation of the unknown mean of a stationary process with a convex correlation function.

(Technically we have replaced the minimization of the generalized variance of the conditioned process, infinite, by the maximization of the generalized variance of the observations, finite, for making comparisons of different designs. To add rigor to this switch requires passing to the limit through finite sets which become dense in $[-1,1]$ - an uninteresting task, and avoided.)

2b) Here is a second example due to Hajek in the same context. For simplicity, let $R(s,t) = (1-|s-t|)_+$ on the interval $[-A,A]$, where $A = m+\alpha$, $0 \le \alpha < 1$. Determine ξ^* through $\xi^*(0) = \xi^*(A) = (m+1)v$, $\xi^*(1) = \xi^*(A-1) = mv, \ldots, \xi^*(m) = \xi^*(A-m) = v$, where $v^{-1} = (m+1)(m+2)$, and compute directly that H^* is constant on $[-A,A]$. Thus the D-optimum design is known in the approximate theory. (When A is an integer, the D-optimum design turns out to be uniform on $0,1,\ldots,A$.) Note further that the designs listed are also G-optimum.

2c) Correlation functions at (5) arise by averaging over suitably scaled product correlation functions. If one does the same with the correlation functions in a) and b) above, the following families surface:

$$R(s,t) = \int \exp\{-\theta|s-t|\}dF(\theta) = \int E_\theta(s,t)dF(\theta); \tag{6}$$

$$R(s,t) = \int (1-\theta|s-t|)_+ dF(\theta) = \int L_\theta(s,t)dF(\theta). \tag{7}$$

The class (7) consists of all convex (or Polya) correlation functions, and (6) is the subclass of completely monotone correlation functions. It turns out that designs that do well for the D-criterion at the correlations that generate these classes, E_θ and L_θ, will have decent efficiency over the full class.

To see this, take the efficiency of a given measure ξ_0 as

$$D - \mathrm{eff}_R(\xi_0) = \min_\xi \iint R(s,t)d\xi(s)d\xi(t) \Big/ \iint R(s,t)d\xi_0(s)d\xi_0(t).$$

Then, for example, if $R(s,t) = \int \exp\{-\theta|s-t|\}dF(\theta) = \int E_\theta(s,t)$ $dF(\theta)$,

$$\begin{aligned} &D - \mathrm{eff}_R(\xi_0) \\ &= \min_\xi \iiint E_\theta(s,t)dF(\theta)d\xi(s)d\xi(t) \Big/ \iiint E_\theta(s,t)dF(\theta)d\xi_0(s)d\xi_0(t) \\ &\geq \int dF(\theta) \min_\xi \iint E_\theta(s,t)d\xi(s)d\xi(t) \Big/ \int dF(\theta) \iint E_\theta(s,t)d\xi_0(s)d\xi_0(t). \end{aligned}$$

This last is at least e^* provided the efficiency at the E_θ's is at least e^* for all θ.

The foregoing argument is given in a more general form in Ylvisaker (1964), where the interest was in estimating mean parameters rather than in locating designs. Changing over to design, one calculation in that paper has D-eff at the uniform measure to be at least 3/4 when R is a convex correlation function, and at least about 7/8 when R is completely monotone. Of course the same device could be used to give lower bounds on efficiency for other measures, such as those which are D-optimum for E_1 or L_1, say.

Going further, there is no difficulty in doing the same with regard to A-efficiencies. We do not trouble to set this out in detail, but we note as an example that if π is the uniform measure on $[-1,1]$, the A-efficiency of the uniform measure over the class of completely monotone function is at least about .872. To check this, set up an efficiency ratio as above and argue that the worst case in the class occurs at some E_θ. Technically then, all that is required is a little calculus, and enough patience to evaluate the minimum of the function $\left[1 - \left(1 - e^{-2\theta}\right)/2\theta\right] / \left(1 - e^{-\theta}\right)$ for $\theta > 0$.

Setting 3: Suppose now that T is some product set $T_1 \times T_2$ and that the covariance kernel of X has a product form, say

$$EX(s,\sigma)X(t,\tau) = R((s,\sigma),(t,\tau)) = R_1(s,t)R_2(\sigma,\tau).$$

We take the case of constant variance so that one reaches the optimality problems at (4) rather than those at (3).

If one determines that ξ_i^* is D-optimum for R_i on the set T_i, $i = 1,2$, then it follows that the product measure $\xi^* = \xi_1^* \times \xi_2^*$ is D-optimum for R on T, as noted in MSY. One merely argues that $H^*(s,\sigma) = H_1^*(s)H_2^*(\sigma)$, so if H_i^* achieves its minimum on the support of ξ_i^*, $i = 1,2$, then H^* achieves its minimum on the support of ξ^*.

This fact has trivially shown up in the first of our settings where T is already a product set. More interestingly, we can construct D-optimum designs as product designs keying off information from the second setting. Thus if

$$R((s,\sigma),\ (t,\tau)) = (1-|s-t|)_+(1-|\sigma-\tau|)_+$$

on $[0,3]\times[0,4]$, the 20 point D-optimum design sits on the full grid of points with integer coordinates. This is an unusual finding for designs in more than one dimension, since one generally expects rectangular grids to be inefficient, Ylvisaker (1975) gives an early example of that. In any event, it is likely the case that the effect of the 20 point design noted above can be captured fairly well by a more cleverly placed 12 point design, say.

Setting 4: Take $T = [-1,1]$ and consider a finite-dimensional process as follows. Let $X(t)$ be a random polynomial, without the zero order term already accounted for by (1), say $X(t) = \Sigma\beta_j t^j$ where $\mathrm{var}(\beta_j) = v_j$, $j = 1,\ldots,k$. Thus $R(s,t) = \Sigma v_j s^j t^j$.

For D-optimality a standard argument says one need only consider ξ symmetric about the origin. Then one finds $H(s) = \Sigma v_{2j}s^{2j}\mu_{2j}$, where μ_{2j} is the $2j^{\text{th}}$ moment of ξ, and insists that ξ^* be supported on the minimum of $2\ \lambda H^*(s) - \Sigma v_j s^{2j}$.

In order to not stray into tedious computations at this point, the quadratic polynomial case is used as an illustration. Thus we seek a measure ξ^* for which the function $2\ \lambda H^*(s) - \Sigma v_j s^{2j} = (2\ \lambda v_2\mu_2 - v_1)s^2 - v_2 s^4$ achieves its minimum on the support of ξ^*. Observe first that if $2\ \lambda v_2 - v_1$ is negative (and recall that $\lambda < 1$) then the function is decreasing in s^2. Thus the measure ξ^* that concentrates on ± 1 and has $\mu_2 = 1$ is D-optimum.

Designs that push toward the boundary of the region are commonplace in the case of D-optimality. A look at the form of the function whose minimum is under scrutiny will convince one that the only alternative to ξ^* here is one which assigns mass to 0 as well as ± 1. Let mass α go on the origin and look to find the minimum of $(2\lambda v_2(1-\alpha) - v_1)s^2 - v_2 s^4$ being attained at 0, ± 1. Of course the value at 0 is 0 and thus one requires that $2\lambda v_2(1-\alpha) - v_1 = v_2$. Regarding λ, v_1 and v_2 as fixed, $(1-\alpha) = (v_1 + v_2)/2\lambda v_2$ should be smaller than 1. Thus $v_1 < (2\lambda - 1)v_2$ and, since $\lambda < 1$, one requires at least that the variance of the quadratic term exceed the variance of the linear term - a sensible finding, though hardly unexpected.

3. Some Remarks.

As mentioned in MSY the asymptotic approach to the Bayesian design problems, as described here, came in response to the difficulty of answering the most basic of questions: are there circumstances in which one should replicate when one has, at best, sparse observation over a large design space? The success of the approach, if any, has been in the production of problems of cleaner appearance and, hopefully, some added insight has been found through them.

The total focus of the paper has been on finding designs, the position is taken that analysis is the easier problem. In particular if a prior is available, Bayesian analysis is straightforward. That the asymptotics employed appear useful in design output is argued through some examples in MSY.

If there is a common thread that runs through the designs found in the various settings of Section 2, it is that one should rely on uniform designs. This lesson is already well known, through there is considerable latitude in implementation of such a principle. The extent to which one turns away from it should depend on specific external information, and the (faithful) modelling of such knowledge.

References

Atkinson, A. and Fedorov, V. V. (1989). Optimum design of experiments. *Encyclopedia of Statistics*, Suppl. Volume, 107-114. John Wiley, New York.

Bandemer, H., Nather, W. and Pilz, J. (1987). Once more: optimal experimental design for regression models (with discussion). *Statistics*, **18**, 171-217.

Cheng, C. S. (1978). Optimality of certain asymmetrical experimental designs. *Ann. Statist.,* **6**, 1239-1261.

Dette, Holger and Studden, William J. (1997). *The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis.* John Wiley, New York.

Fedorov, V. V. (1972). *Theory of Optimal Experiments* (Translated and Edited by E. M. Kleinko and W. J. Studden). Academic Press, New York.

Hajek, J. (1956). Linear estimation of the mean value of a stationary random process with convex correlation function. *Czech. Math. J.* (Also in *Selected Translations in Math. Statist. and Prob.,* **2**, 41-61.)

Hajek, J. (1962). An inequality concerning random linear functionals on a linear space with a random norm and its statistical applications. *Czech. Math. J.,* **12**, 486-491.

Hardin, R. H. and Sloane, N. J. A. (1993). A new approach to the construction of optimal designs. *J. Statistical Planning and Inf.,* **37**, 339-369.

Johnson, M. E., Moore, L. M. and Ylvisaker, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference*, **26**, 131-148.

Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Ann. Math. Statist.,* **29**, 675-699.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *J. Royal Statist. Soc.* Series B, **21** 272-319.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.,* **30**, 271-294.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.,* **12**, 363-366.

Mitchell, T. J., Morris, M. D. and Ylvisaker, D. (1995). Two-level fractional factorials and Bayesian prediction. *Statistica Sinica*, **5**, 559-573.

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

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *J. Roy. Statist. Soc.* Ser. B, **40**, 1-42.

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

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

Sacks, J. and Ylvisaker, D. (1985). Model robust design in regression: Bayes theory. *Proceeding of the Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (LeCam and Olshen, Eds.) **2**, 667-679. Wadsworth, Monterey, CA.

Steinberg, David M. and Hunter, William G. (1984). Experimental design: review and comment (with discussion). *Technometrics*, **26**, 71-130.

Ylvisaker, D. (1964). Lower bounds for minimum covariance matrices. *Ann. Math. Statist.,* **35**, 362-368.

Ylvisaker, D. (1975). Designs on random fields. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.), 593-607. North Holland, Amsterdam.

Ylvisaker, D. (1987). Prediction and design. *Ann. Statist.,* **15**, 1-19.