Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Maximum entropy sampling

M. C. Shewry [a] & H. P. Wynn [a]
[a] The City University , London
Published online: 28 Jul 2006.

To cite this article: M. C. Shewry & H. P. Wynn (1987) Maximum entropy sampling, Journal of Applied Statistics, 14:2, 165-170, DOI: 10.1080/02664768700000020

To link to this article: http://dx.doi.org/10.1080/02664768700000020

Journal of Applied Statistics, Vol. 14, No. 2, 1987

Maximum entropy sampling

M. C. SHEWRY & H. P. WYNN, *The City University, London*

1 Introduction

There are several distinct strands of research in the history of optimum or efficient data collection. Within the area of experimental design the British School started with the work of R. A. Fisher & F. Yates in agricultural field trials and led to a huge literature in combinatorial design. The American decision theoretic school developed out of the foundations of statistical design theory laid down by A. Wald. Within this, two rather distinct strands emerged: the theory of optimum design of J. Wolfowitz & J. Kiefer and the information theory approach of D. Blackwell, L. Le Cam and D. V. Lindley. These have been studied recently under the banner of Bayesian design of experiments (see Chaloner, 1984, for detailed discussion).

The wider field of data collection must include survey sampling. Spatial sampling can be considered as a merging of ideas from survey sampling and optimum design. A recent article which draws together many of the ideas is Ylvisaker (1986).

In the present paper a simple criterion and examples are presented in a deliberate attempt to understand the spectrum between controlled experiment and more passive observation. This contrast has been discussed with some philosophical references in Wynn (1982a). The other distinction which we hope to illuminate is that between observation on a finite, closed, system and observation in an infinite or wider environment.

2 Information and entropy

We take as our starting point the classical idea of 'the amount of information in an experiment' due to Blackwell (1951) and developed by Lindley (1956) from a Bayesian point of view.

Let e be an experiment on a random system X with sampling density $p(x|\theta)$ and parameter θ. If the prior distribution for θ is $\pi(\theta)$ then if x is the data obtained on performing e the posterior density for θ is

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta).$$

The amount of information on θ, before e, is

$$I(\theta)=\int \pi(\theta)\log\, \pi(\theta)d\theta$$

and after the experiment is

$$I(\theta|x)=\int \pi(\theta|x)\log\, \pi(\theta|x)\mathrm{d}\theta.$$

Thus the increase in information is

$$I(\theta|x)-I(\theta) \tag{1}$$

In comparing experiments, *a priori*, we wish to consider the value of the expected increase, that is the expected value of (1) over the full joint distribution of X and θ. We can write this as

$$\begin{aligned} g(e) &= \text{Expected value of (1)} \\ &= E_x E_\theta \log\left[\frac{\pi(\theta|x)}{\pi(\theta)}\right]. \end{aligned}$$

The entropy of the system with respect to θ is defined as $\mathrm{Ent}(\theta)=-I(\theta)$.

Ideally one may select an experiment with the maximum value of $g(e)$ for e in E. Several difficulties may arise. The choice set may be too large so that we are comparing experiments in rather different environments and the criterion is too coarse. Another difficulty is that what constitutes a parameter of interest may change. One may not so much be interested in an internal parameter of the model as the future behaviour of the process. Although the Bayes methodology makes no distinction there are serious difficulties as to the definition of the full system. Should one consider future observations as embedded in a larger, possibly infinite, system?

3 Finite systems

A resolution of the problems of defining E is to consider a finite system and experimentation as a glimpse at this finite system. This seems very reasonable for observation on either a finite population, as occurs in survey sampling, or on a finite machine or component in an engineering context. Finite here means that the total number of potential observation sites is finite.

Thus let $S=\{1,\ldots, N\}$ be a finite system which is described by a random vector $(Y_1,\ldots, Y_N)$ where we consider Y_i attached to a unit (observation site) $i \in S$. Partition S into two disjoint sets the sample s and its complement $\bar{s}$: $s \cup \bar{s}=S$ and $s \cap \bar{s}=\phi$. *Let* $\#(s)=n$, which we call the sample size. We consider all the Y_i, $i \in s$, the sample, as observed. The full joint distribution of the Y_i, $i \in S$ is known and the statistical problem is to interpolate for Y_i, $i \in \bar{s}$, from Y_i, $i \in s$. Let Y_s and $Y_{\bar{s}}$ be the vectors of sampled and unsampled Y_i.

We now use as our criterion the analogue of $-g(e)$ from Section 1. Let $Y_{\bar{s}}|Y_s$ denote the random variable $Y_{\bar{s}}$ conditional on Y_s for a random variable X with density $f(X)$ we define the entropy as

$$\begin{aligned} \mathrm{Ent}(X) &= -E_X(\log f(X)) \\ &= -\int f(x)\, \log f(x)\mathrm{d}x. \end{aligned}$$

The critical property of Ent(.) for S which derives from the classical results on information, is

$$\text{Ent}(Y_S)=\text{Ent}(Y_s)+E_{Y_s}\{\text{Ent}(Y_{\bar{s}}|\,Y_s)\} \tag{2}$$

Here Y_S refers to the whole population S and the expectation is with respect to the marginal (unconditional) distribution of Y_s. We assume that all expectations exist.

The second term on the right hand side of (2) is just the $-g(e)$ of Section (1) with $Y_{\bar{s}}$ playing the role of θ and e the choice of sample, s. The optimum design problem is to minimise the second term. Then since $\text{Ent}(Y_S)$ is fixed and finite this problem is equivalent to

$$\text{maximise}_{s}\ \text{Ent}(Y_s),$$

hence the term 'maximum entropy sampling'.

This criterion is different in philosophy from those which arise in classical optimum design. Here we wish to absorb into the sample the maximum amount of variability so that *conditionally* on the sample the unsampled population has minimum variability.

The finiteness of S is critical. In more standard settings one would not at first glance normally seek to design in such a way that the observed part had minimal *internal* information. But reflection about the proper conditional statement might indeed lead to something closer to the present prescription. We note that it is the *fixed* nature of $\text{Ent}(Y_S)$ that is special here and that $s \subseteq S$. If there are several populations $S_1, \ldots, S_L$ and samples $s_i \subseteq S_i$ $(i-1, \ldots, L)$ and different interpolation problems for $\bar{s}_i=S_i \setminus s_i$ then the choice set becomes more problematical. Each S_i may be qualitatively different and the system may not be closed in the same sense.

4 The Gaussian case

Let $Y_S=[Y_s\ Y_{\bar{s}}]$ have a joint multivariate normal distribution with covariate matrix (with obvious notation)

$$\Gamma_s=\begin{bmatrix}\Gamma_s & \Gamma_{s,\bar{s}}\\ \Gamma_{\bar{s},s} & \Gamma_{\bar{s}}\end{bmatrix}$$

In this case

$$\text{Ent}(Y_S)=\text{constant}+\log\det(\Gamma_S).$$

Except for constants the entropy expansion is

$$\log\det(\Gamma_S)=\log\det(\Gamma_s)+\log\det(\Gamma_{\bar{s}}-\Gamma_{\bar{s},s}\Gamma_s^{-1}\Gamma_{s,\bar{s}}).$$

The second matrix on the right hand side is the conditional covariance matrix for $Y_{\bar{s}}|\,Y_s$ and we write it $\Gamma_{\bar{s}/s}$. Thus the criterion becomes simply

$$\max\log\det(\Gamma_s)$$

or simply $\max\det(\Gamma_s)$. This conversion of a conditional problem to an unconditional problem is of great computational benefit. Other criteria are discussed by Sacks & Schiller (1986) in parallel work. Of particular interest is the minimax criterion to minimise the worst extrapolation error over $\bar{s}$. This is to minimise over s the maximum diagonal term of $\Gamma_{\bar{s}|s}$. For some very restricted models maximum entropy and the latter criterion coincide (this will be the subject of a joint paper with J. Sacks). A

further class of criteria is to maximise trace $(\Gamma_{s|s}^{p})$ $(p>0)$, but none of these seems to have the attractive decomposition afforded by entropy, except in special cases.

4 Implementation

We discuss below two examples using max det (Γ_s). Both are amenable to a simple exchange algorithm in which a single point in s is exchanged for a point in $\bar{s}$. Extension to multiple exchange or 'excursion' algorithms following the ideas of Mitchell (1974) and others is under way. This includes close collaboration with J. Sacks at the University of Illinois, Urbana-Champaign, and co-workers, who have carried out extensive work with the simulated annealing algorithm. It is typical of these computational problems in spatial sampling that local optima are prevalent and experience is needed with combinations of exchange and stochastic relaxation methods.

In both examples below we take a 101×101 grid of $(0,\ldots,100)^2$ so that $N=101^2$ and $n=20$. The solutions are based on up to 80,000 single point exchanges.

Example 1 Linear model with stationary error

We consider a standard linear model

$$Y_S=X_S\theta+\varepsilon$$

where $E(\varepsilon)=0$, $\text{var}(\varepsilon)=\tilde{\Gamma}_S$ and θ is $p\times1$. We can make this into a Bayes model by giving θ a suitable prior distribution. For *simplicity* take the prior covariance $\text{var}(\theta)=\sigma^2 I_p$. Then

$$\Gamma_S=\tilde{\Gamma}_S+\sigma^2 X_S X_S^T$$

and similarly for the sample

$$\Gamma_s=\tilde{\Gamma}_s+\sigma^2 X_s X_s^T,$$

with obvious notation.
Using a standard formula

$$\det(\Gamma_s)=(\det\tilde{\Gamma}_s)\ (I_P+\sigma^2 X_s^T\,\tilde{\Gamma}_s^{-1}X_s).$$

for stationary error we can write $\tilde{\Gamma}_s=\sigma_\varepsilon^2 V_s$ where V_s is a correlation matrix. Then

$$\det(\Gamma_s)=(\sigma_\varepsilon^2)^p\ \det(V_s)\det(I_p+\alpha X_s^T V_s^{-1}X_s)$$

where

$$\alpha=\frac{\sigma^2}{\sigma_g^2}.$$

This may be interpreted as a 'signal-to-noise' ratio or variance ratio by analogy with random effect models. As α increases we can watch the effect on the optimum sampling. As $\alpha\to\infty$ so that the model dominates we tend to a singular covariance model since typically $p\ll n$. However, for finite α we have a balance between the model and the error which persists as $\alpha\to\infty$. For large α the model term behaves like $\det(X_s^T V_s^{-1}X_s)$ which is D-optimality for θ in a Bayesian framework. The solution will tend to push the observation towards the extremes. When $\alpha=0$ however we have pure noise and if Γ_S arises from a stationary process in space observations will be 'passively' spread over the design region.

We consider a two dimensional example:

$$E(Y_{x_1,x_2}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \varepsilon.$$

and if ε_1 and ε_2 are the errors at $x^{(1)}$ and $x^{(2)}$ then

$$\text{cov}(\varepsilon_1, \varepsilon_2) = \varepsilon^{-\lambda d_{12}}$$

where $d_{12} = \|x^{(1)} - x^{(2)}\|$. The full covariance matrix is

$$\Gamma_s = \{e^{-\lambda d_{ij}} + \sigma^2(x^{(i)T}x^{(j)} + 1)\}.$$

Since $e^{-\lambda d12} \to 1$ as $x^{(1)} \to x^{(2)}$ we have $\alpha = \sigma^2$. The results are given for various values of λ and α. The x-values are centred on (50, 50).

Fig. 1 shows that as the model term dominates 'holes' start appearing. This leads to a strong suspicion of a 'phase change' effect with observations playing the role of particles in statistical mechanics. The existence of 'holes' in the pure D-optimum case is evident from the work on constrained D-optimality (Wynn, 1982b). Indeed, as $\alpha \to \infty$ and $\lambda \to \infty$ we tend to the D-optimum solution since $V_s \to I_p$. A theory is needed in which observations become *dense* over the design region so that the critical change can be investigated.

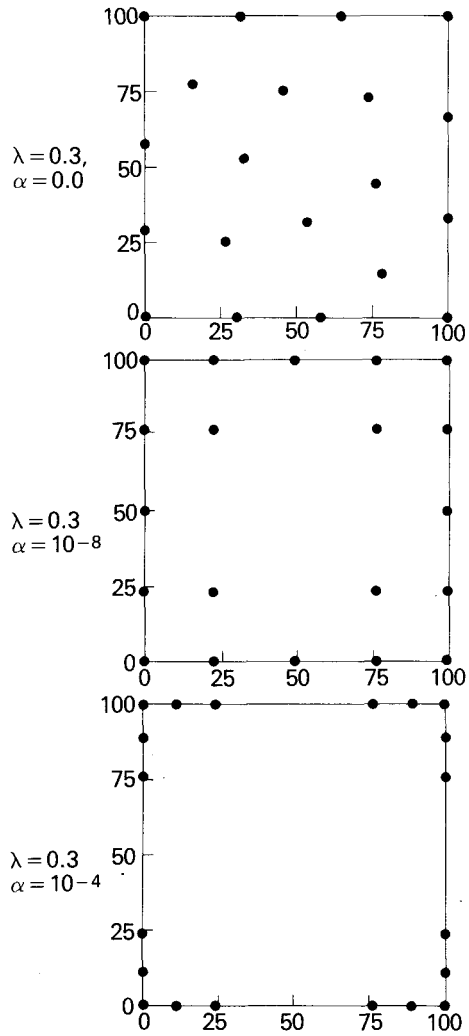

FIG. 1. Optimum sampling, linear model, stationary error.

Example 2 Brownian Sheet

This is an example of a pure covariance process which is non-stationary. The sampling problem is discussed in Ylvisaker (1975). The process is defined on R_2^+. It has zero mean and the covariance between $Y_{x(1)}$ and $Y_{x(2)}$ where $x^{(1)}=(x_{11}, x_{12})$ and $x^{(2)}=(x_{21}, x_{22})$ is $\min(x_{11}, x_{21})\min(x_{12}, x_{22})$. Since the process is tied down on the x_1 and x_2 axes we expect that sampling over a square would tend to push observations towards the top right hand corner. This is evident from Fig. 2.

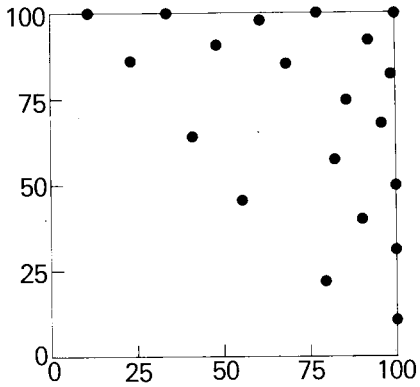

FIG. 2. Optimum sampling, Brownian sheet.

Acknowledgements

We are grateful to Jerry Sacks and to our colleague Russell Gerrard for helpful suggestions. Toby Mitchell inspired this work and the Gaussian case should be credited to him.

Correspondence: Professor H. P. Wynn, University Statistical Laboratory, Department of Mathematics, The City University, Northampton Square, London EC1V 0HB, UK.

REFERENCES

BLACKWELL, D. (1951) Comparison of experiments, *Proceedings of the 2nd Berkeley Symposium*, pp. 93–102 (Berkeley, Ca., University of California Press).

CHALONER, K. (1984) Optimal Bayesian design for linear models, *Annals of Statistics*, 12, pp. 283–300.

LINDLEY, D.V. (1956) On a measure of information provided by an experiment, *Annals of Mathematical Statistics*, 27, pp. 986–1005.

MITCHELL, T.J. (1974) Computer generations of D-optimal first-order designs, *Technometrics*, 16, pp. 211–220.

SACKS, J. & SCHILLER, S. (1986) Spatial designs, in: GUPTA, S.S. (Ed.) *Fourth Purdue Symposium on Statistical Decision Theory and Related Topics* (London, Academic Press).

YLVISAKER, D. (1975) Designs on random fields, *A Survey on Statistical Design and Linear Models*, pp. 593–607 (Amsterdam, North Holland).

YLISAKER, D. (1987) Prediction and design, *Annals of Statistics* (to appear).

WYNN, H.P. (1982a) Controlled versus random experimentation, *The Statistician*, 31, pp. 237–244.

WYNN, H.P. (1982b) Optimum submeasures with application to finite population sampling, in: GUPTA, S.S. (Ed.) *Third Purdue Symposium on Statistical Decision Theory and Related Topics* (London, Academic Press).