Bayesian Experimental Design and Shannon Information

Paola Sebastiani, City University and **Henry P. Wynn, University of Warwick**

Abstract

The information theoretic approach to optimal design of experiments yields a simple design criterion: the optimal design minimizes the expected posterior entropy of the parameters. Unfortunately, this strategy is often computational infeasible for non-linear problems and numerical approximations are required. This paper reviews the information theoretic approach to design of experiments, and examines computational issues related to the minimization of the expected posterior entropy and its asymptotic approximations. It is shown that Maximum Entropy Sampling simplifies both the formulation of the design criterion and the optimization problem. Numerical advantages are shown in an example, where the exact solution is compared to asymptotic optimal solution.

Keywords: Entropy, Optimal Experimental Design, Integration, Maximum Entropy Sampling..

Address: Paola Sebastiani, Department of Actuarial Science and Statistics, City University, London EC1V 0HB, United Kingdom, Northampton Square, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (171) 477-8959, FAX: +44 (171) 477-8838,

EMAIL: p.sebastiani@city.ac.uk,

URL: http://www.city.ac.uk/~sn303.

1. Introduction

Suppose we have a set of possible experiments denoted by Ξ. The outcome of the experiment is a random n-vector Y in $\mathcal{Y}$, and we assume that, given $\Theta = \theta$ and an experiment ξ, Y has a known distribution with density $p(y|\theta, \xi)$. The parameter vector Θ has prior density $p(\theta)$ not dependent on the experiment ξ. The problem considered in this paper is the optimal choice of ξ from the set of possible experiments Ξ so that the maximum amount of information about Θ is obtained, when the amount of information is measured as the negative Shannon entropy. Let Γ be a N-random vector, with p.d.f $p(\gamma)$. The Shannon entropy is defined as

$$\mathrm{Ent}(\Gamma) = E_\Gamma\{-\log p(\Gamma)\}.$$

The Bayesian information theoretic approach to choice of experiment was originally proposed by Lindley (1956) and consists of choosing the experiment that minimizes the expected posterior (Shannon) entropy of Θ. The argument goes as follows. The expected gain of information from ξ is

$$I = \mathrm{Ent}(\Theta) - E_Y \mathrm{Ent}(\Theta|Y, \xi), \tag{1}$$

where the expectation is over the marginal distribution of Y. If $\mathrm{Ent}(\Theta)$ is not design dependent, which is an assumption made throughout this paper, maximization of (1) is equivalent to finding ξ in Ξ that minimizes the overall expected risk

$$E_Y\{\mathrm{Ent}(\Theta|Y, \xi)\}. \tag{2}$$

Note that this can also be shown to be the coherent solution of a decision problem with loss function the log-score introduced by Good (1952) in the context of probability assessment.

Thus $e(\xi) = E_Y\{\mathrm{Ent}(\Theta|Y, \xi)\}$ is the design criterion. Given the well known computational problems of Bayesian analysis, the feasibility of this approach has been limited to simple problems, in which, conditional on Θ, the sampling distribution of Y is normal and the expectation is a linear function of the parameters. In general, for non-linear/non-normal problems, the posterior distribution of Θ cannot be found in closed form, so that the formulation of the design criterion is not possible. Common solution is to rely on an asymptotic approximation of the posterior distribution from which an asymptotic design criterion is derived. This clearly leaves open the problem of the optimal solution of design problems along this information theoretic approach when the sample size allowed for the experiment is small. Needless to say that this is the case in which an optimal solution is mostly required.

In this paper we show that, in some cases, there exists a dual solution to the design problem that partly overcomes computational difficulties.

The approach is due to Sebastiani and Wynn (1997) and extends a method described in Shewry and Wynn (1987) for predictive design problems to estimative design problems. The main result is that, under general assumptions, the minimization of the expected posterior entropy of Θ w.r.t. the experiment is achieved by maximizing the marginal entropy of the observations. This yields an alternative design criterion called *Maximum Entropy Sampling* (MES). In Section 2, we describe the theory that yields the formulation of the dual design criterion. Computational advantages are discussed in Section 3, and an example is given in Section 4

2. Theory

Sebastiani and Wynn (1997), introduce MES, for estimation problems. The principle is that if the entropy of the sampling distribution is not functionally dependent on the design, then a design minimizing the expected posterior entropy can be found by maximizing the marginal entropy of the data. The result is obtained by using an identity which appears in several forms in information theory, for example Cover and Thomas (1991). Suppose that the random N-vector Γ is decomposed in $(\Gamma_s : \Gamma_{\bar{s}})$ where $s \in \{1, \ldots, N\}$. Then

$$\mathrm{Ent}(\Gamma) = \mathrm{Ent}(\Gamma_s) + E_{\Gamma_s}\{\mathrm{Ent}(\Gamma_{\bar{s}}|\Gamma_s)\}.$$

Now let $\Gamma = (\Theta, Y)$ and apply the identity above twice, by interchanging the role of Θ and Y:

$$\begin{aligned} \mathrm{Ent}(\Theta, Y) &= \mathrm{Ent}(Y) + E_Y\{\mathrm{Ent}(\Theta|Y)\} && (3) \\ &= \mathrm{Ent}(\Theta) + E_\Theta\{\mathrm{Ent}(Y|\Theta)\}. && (4) \end{aligned}$$

From (3) we see that, if $\mathrm{Ent}(\Theta, Y)$ is fixed before the experiment, then minimizing $E_Y\{\mathrm{Ent}(\Theta|Y)\}$ w.r.t. ξ is equivalent to maximizing $\mathrm{Ent}(Y)$. From (4), we derive a sufficient condition for $\mathrm{Ent}(\Theta, Y)$ be fixed. Since we are assuming that the prior distribution of Θ is not functionally dependent on the experiment, then $\mathrm{Ent}(\Theta, Y)$ will not be functionally dependent on ξ as long as $\mathrm{Ent}(Y|\Theta)$ is not functionally dependent on ξ. Thus, the MES principle can be applied, for instance, to regression type experiments, for sampling models with additive errors, and variance which is not design dependent:

$$Y|(\xi, \theta) = \mu(\xi, \theta) + \epsilon.$$

Here μ is the fixed model, conditional on $\Theta = \theta$, and ϵ is a vector of i.i.d. errors independent of θ and ξ. The experiment is $\xi = \{x_1, \cdots, x_n\}$, where x_i are control variables that can be chosen by the experimenter in a design region $\mathcal{X}$, and the sample size n allowed for the experiment is fixed. Given θ, $\mu(\xi, \theta)$ is fixed so that

$$\mathrm{Ent}(Y|\xi,\theta) = \mathrm{Ent}(\epsilon)$$

and hence $\mathrm{Ent}(Y|\xi,\theta)$ is not functionally dependent on the design. A key aspect is that $\mu(\theta,\xi)$ can be any function of ξ and θ and many non-linear regression models fall into this framework.

In Sebastiani and Wynn (1997), MES designs are found for the particular case of a discrete prior distribution on the parameters of a normal model. It is shown that when the prior distribution is concentrated on two points, θ_1 and θ_2, then the optimal design has one support point x^* that maximizes the distance between the two conditional means: $(\mu(x,\theta_1)-\mu(x,\theta_2))^2$. This result is independent of the prior weights. For the more general case of a discrete prior supported on k points, an approximation of the entropy is provided, that can be used when the sampling variance becomes large. Furthermore, heuristic arguments show that the upper bound on the support size of the MES design is $k(k-1)/2+1$.

3. Computation

The usual approach to design of experiment for non-linear problems relies on asymptotic approximations of the posterior distribution of Θ. Suppose that Θ has a continuous prior, and that both the sampling and the prior density are positive and twice differentiable near the Maximum Likelihood Estimate (MLE) $\hat{\theta}$ of θ. Several *normal* approximations of the posterior distribution are possible, (Berger, 1985, page 224). Most familiar ones are those based either on the observed or the expected Fisher information matrix as approximation of the inverse of the posterior dispersion.

Let $l(\theta,y)$ be the log-likelihood function. The Fisher information matrix is defined as:

$$I(\theta,y,\xi) = -\left(\frac{\partial^2 l(\theta,y)}{\partial\theta_i\partial\theta_j}\right).$$

The *observed* Fisher information matrix is then $I(\hat{\theta},y,\xi)$: the Fisher information matrix evaluated in $\hat{\theta}$. This will depend on the data y directly, and via the MLE. The *expected* Fisher information matrix is $I(\theta,\xi)=E_Y\{I(\theta,y,\xi)\}$. The expectation is over the sampling distribution of $Y|\theta,\xi$, so the expected Fisher information matrix is a function of the parameters. The posterior distribution can be then approximated as $\Theta|y,\xi \sim N(\hat{\theta},V_p)$, where V_p is either $I(\hat{\theta},y,\xi)^{-1}$ or $I(\hat{\theta},\xi)^{-1}$. Usually $I(\hat{\theta},y,\xi)$ gives a more refined approximation but, for exponential family models, the two approximations are the same.

It is well known that, if $\Gamma \sim N(\mu,\Sigma)$, then up to a constant, $\mathrm{Ent}(\Gamma) = \log\det\Sigma$. Thus the asymptotic normal distribution of $\Theta|y$ reduces the search of the design to minimizing $E_Y(\log\det V_p)$, where the expectation is over the marginal distribution of Y. When $V_p = I(\hat{\theta}, \xi)^{-1}$, so that the posterior dispersion depends on the data only via $\hat{\theta}$, consistency of MLE allows us to replace $E_Y(\log \det V_p)$ by

$$-E_\Theta\{\log \det I(\theta, \xi)\} \tag{5}$$

where the expectation is now over the prior distribution of Θ, see Chaloner and Verdinelli (1995) for further details and a discussion of other approximations. A design minimizing (5) is called (asymptotic) D-optimal. A first order approximation of (5) about the prior mean θ_0 yields *local* D-optimality: $-\log \det I(\xi, \theta_0)$. Application of this approximation to experimental designs for generalized linear model is given in Sebastiani and Settimi (1997, 1998).

From a computational point of view, the MES principle replaces sometimes difficult computations with conditional densities by computations with the full marginal density of the observations: $p(y|\xi) = \int_\Omega p(y, \theta|\xi)p(\theta)d\theta$. Thus we need to maximize

$$\mathrm{Ent}(Y|\xi) = -\int_{\mathcal{Y}} \log\{p(y|\xi)\}p(y|\xi)dy \tag{6}$$

over the set of possible design Ξ. The evaluation of the entropy involves an integration over the sample space, and the evaluation of the marginal density $p(y|\xi)$ involves an integration over the parameter space, since:

$$p(y|\xi) = \int_\Omega p(y|\theta, \xi)p(\theta)d\theta. \tag{7}$$

Sebastiani and Wynn (1998) discuss integration techniques for a numerical approximation of (6) and they show how, at least numerically, the design problem can be reduced to the one we have when the parameter Θ is a discrete random vector. The objective function can be approximated by evaluating the integrals (6) over a grid of points for Y, say $\mathcal{Y}_{\text{grid}} = \{y_1, \ldots, y_h\}$, with weights $\{w_1, \ldots, w_h\}$, so that (6) is replaced numerically by:

$$-\sum_i w_i p(y_i|\xi) \log p(y_i|\xi).$$

The evaluation of the marginal density $p(y_i|\xi)$ is done by using quadrature formulae so that $p(y_i|\xi)$ is approximated by

$$\sum_j v_j p(\theta_j)p(y_i|\theta_j, \xi) \tag{8}$$

where $\{\theta_1, \ldots, \theta_k\}$ are quadrature points on the parameter space, and $\{v_1, \ldots, v_k\}$ are quadrature weights. This shows that, up to a normalizing constant, (8) is the marginal density $p(y_i|\xi)$ obtained when Θ is a discrete random variable taking values θ_j with prior probability $p(\theta_j v_j)$.

The goodness of the approximation depends on the choice of quadrature points, that can be obtained, for instance, as Cartesian product of one dimensional grids of points. More efficient approximations can be based on integration on Lattices (Sloan & Joe, 1994). The computational efficiency clearly depends on the precision required. Stochastic evaluation of either integrals could also be investigated.

Clearly, the approach is feasible as long as the dimension of $\mathcal{Y}$ is small. As the sample size allowed for the experiment increases, then the minimization of $E\{\mathrm{Ent}(\Theta|Y)\} \approx -E\{\log\det I(\theta,\xi)\}$ will be preferable.

A further approximation can be obtained when the observations $Y = (Y_1,\ldots,Y_n)$ are such that

$$Y_i|\theta = \mu(x_i,\theta) + \epsilon_i$$

and the ϵ_i are i.i.d. $N(0,\sigma^2)$. Consider the numerical approximation of the marginal density of Y. Let $v = \sum_j v_j p(\theta_j)$, and $v'_j = p(\theta_j)v_j/v$, $(j = 1,...,k)$, so that (8) is $v\sum_j v'_j p(y_i|\theta_j,\xi)$. The numerical approximation of the entropy of Y is:

$$-v\log v - v\sum_j v'_j \int_Y p(y|\theta_j,\xi)\log\sum_j p(y|\theta_j,\xi)v'_j dy$$

which compared to the formula of the entropy for a discrete prior in Sebastiani and Wynn (1997) differs by the correction factor $-v\log v$, and the multiplicative factor v. Note that the correction factor is 0 whenever $v = 1$. An approximation using two quadrature points would imply that the optimal design is a one-point design. As the range of values for Θ becomes large, than the number of quadrature points will increase, and this will increase the upper bound on the support size of the MES design: $k(k-1)/2+1$. The dependency of the upper bound on the number of quadrature points explains another phenomenon concerning support size: as the prior information becomes small the support of the optimal design increases (Chaloner & Verdinelli, 1995). When the sampling variance is large enough, we can use the approximation of the entropy found in Sebastiani and Wynn (1997), so that

$$\begin{aligned}\mathrm{Ent}(Y|\xi) &= v\Bigg[-\log v + \frac{n}{2}\{\log(2\pi\sigma^2)+1\} \\ &+ \frac{1}{4\sigma^2}\sum_{j=1}^{k}\sum_{l=1}^{k} v'_j v'_l d_{jl} + O\left(\frac{1}{\sigma^4}\right)\Bigg].\end{aligned} \tag{9}$$

where $d_{jl} = \sum_i(\mu(x_i,\theta_j) - \mu(x_i,\theta_l))^2$. The function (9) is separable in x_i, so the optimal design will be a one point design concentrated in

	Parameter Space			
σ	[0.5,0.8]	[0.5,1]	[0.5,5]	[0.5,10]
0.35	1.37	1.30	0.46	0.26
0.40	1.55	1.37	0.46	0.26
0.45	1.56	1.37	0.46	0.26
0.50	1.56	1.37	0.46	0.26
x_{app}	1.56	1.37	0.46	0.26
x_{asym}	1.54	1.45	1	1

Table 1: Support of the optimal design for different values of the standard deviation σ, and different parameter space. x_{app} is the point that maximizes the approximation in (9), x_{asym} is the support point of the asymptotic optimal design.

$$x^* = \text{argmax}_{\mathcal{X}} \sum_{j=1}^{k} \sum_{l>j}^{k} v'_j v'_l \{\mu(x, \theta_j) - \mu(x, \theta_l)\}^2.$$

Thus, at least numerically, when the sampling variance becomes large, optimal designs concentrate on one point for continuous prior.

4. Example

In this section, we apply the results of section 3 to design the experiment for a first order decay model:

$$\begin{array}{rcc} Y|\theta & \sim & N(\mu, \sigma^2),\ x \in [0,2] \\ \mu(x,\theta) & = & \exp(-\theta x), \\ \Theta & \sim & U(c,d). \end{array}$$

Aim of the example is (i) to examine the sensitivity of the optimal design to different values of the sampling variance σ^2 and parameter space (c,d); (ii) to study the goodness of the approximation given in (9); (iii) to compare the solution found by maximizing the approximation (9) to the solution based on an asymptotic approximation of the posterior distribution of the parameters.

Material and Method Four values of the sampling standard deviation were considered: $\sigma = 0.35, 0.4, 0.45, 0.5$. For each value of σ, four parameter spaces were chosen: [0.5,0.8], [0.5,1], [0.5,5], and [0.5,10], and the best one-, two- and three-point design were found by maximizing the entropy of the marginal distribution of the data. The evaluation of the entropy was based on 20^n Hermite quadrature points, where n is the sample size allowed for the experiment, i.e. $n = 1, 2, 3$, and the marginal density of Y was evaluated

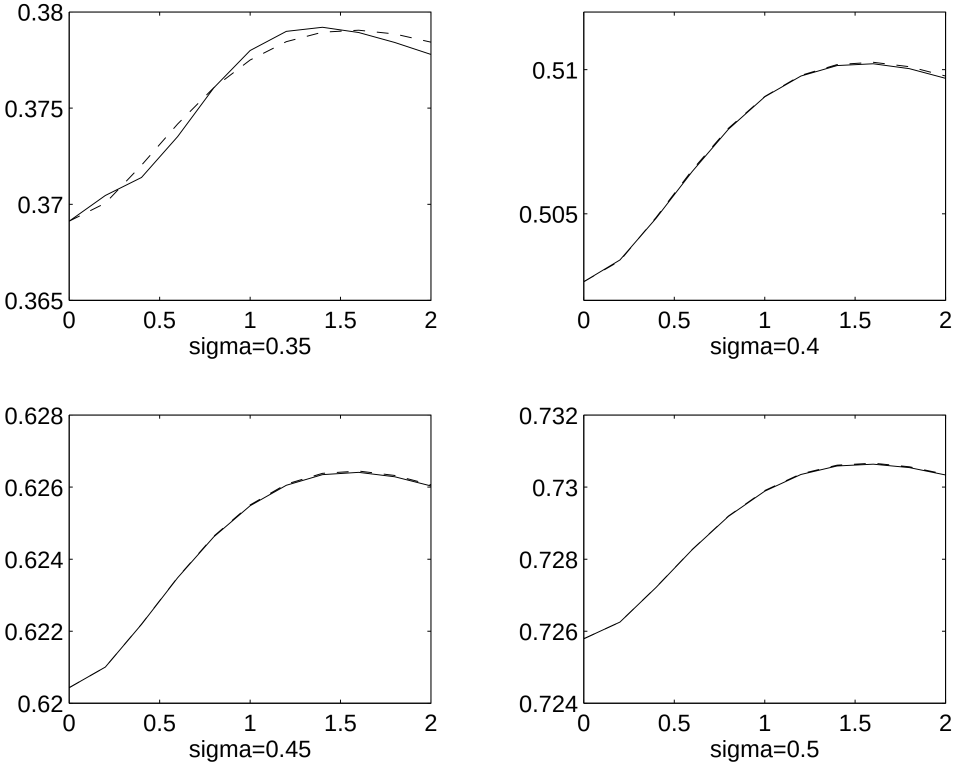

Figure 1: Numerical evaluation of the entropy (continuos line) and approximation using (9): parameter space = [0.5;0.8].

using 20 Legendre quadrature points. The maximization was then performed numerically by using constrained optimization routines available in `Matlab`.

The approximation (9) was evaluated using 20 Legendre quadrature points.

The asymptotic optimal designs were found as follows. Chaloner (1993) shows that if Θ has a continuous prior supported on $[c, d]$ such that $[1/d, 1/c] \subset [0, 2]$ and $d/c \leq (2 + \sqrt{2})/2$, then the asymptotic optimal design is a one point design supported at $E(\Theta)^{-1}$. The same design is locally optimal, if $E(\Theta)$ is the best guess of Θ (Kitsos, 1995). Of the four parameter spaces considered, only [0.5,0.8] satisfies Chaloner's condition, since $[1/d, 1/c] = [1.25, 2]$ and $.8/.5 = 1.6 \leq (2+\sqrt{2})/2$. Thus the asymptotic optimal design is supported on $E(\Theta)^{-1} = 1.54$. For larger values of d, the asymptotic optimal design was found by minimizing (5). The expected Fisher information of the n-point design $\xi = \{x_1, \ldots, x_n\}$ is $\sum_i x_i^2 \exp(-2\theta x_i)$. Thus, the asymptotic optimality criterion

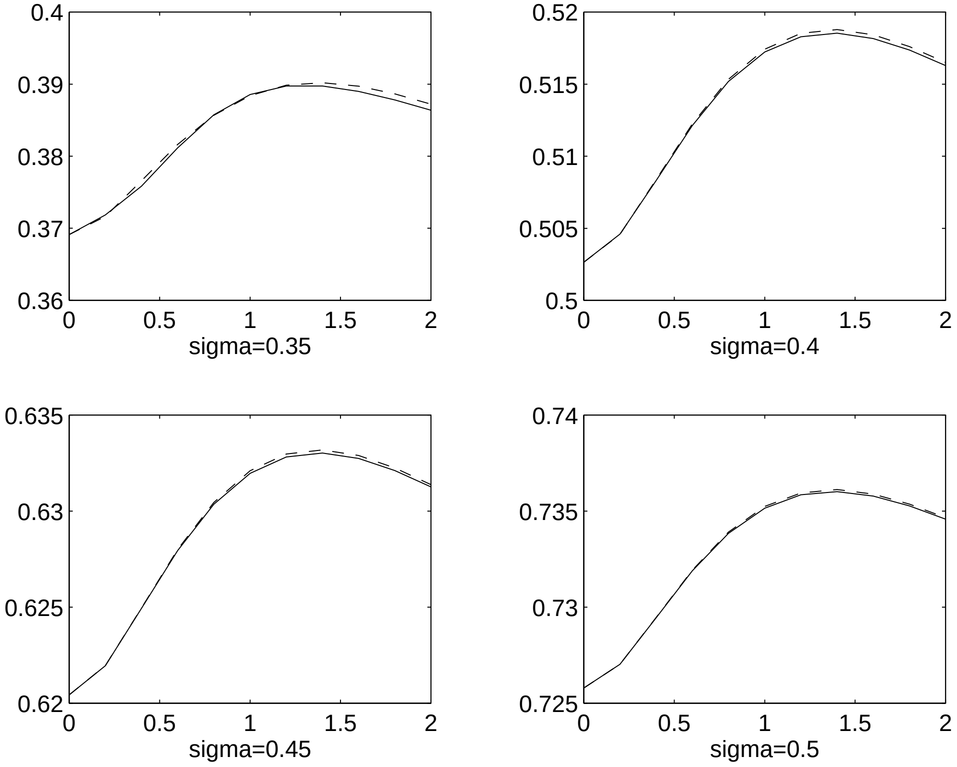

Figure 2: Numerical evaluation of the entropy (continuos line) and approximation using (9): parameter space = = [0.5,1].

$$\sum_i \int_c^d x_i^2 \exp(-2\theta x_i)/(d-c)d\theta$$

was evaluated by numerical integration using 20 Legendre integration points, and the maximization was performed numerically by using constrained optimization routines available in `Matlab`.

Results In all cases considered a one point design was optimal, and the support points are given in Table 1 for the 12 combinations of values for σ^2 and parameter space. The optimal support point that maximizes the approximation (9) is extremely close to the optimal solution, and the goodness of the approximation can be seen in Figures 1, 2, 3 and 4.

The asymptotic optimal design (last row of Table 1) gives a good approximation of the exact solution when the parameter space is small, but as soon as the parameter space becomes large, the solution is far from the exact one. Clearly, as the prior precision is small, more observations are needed

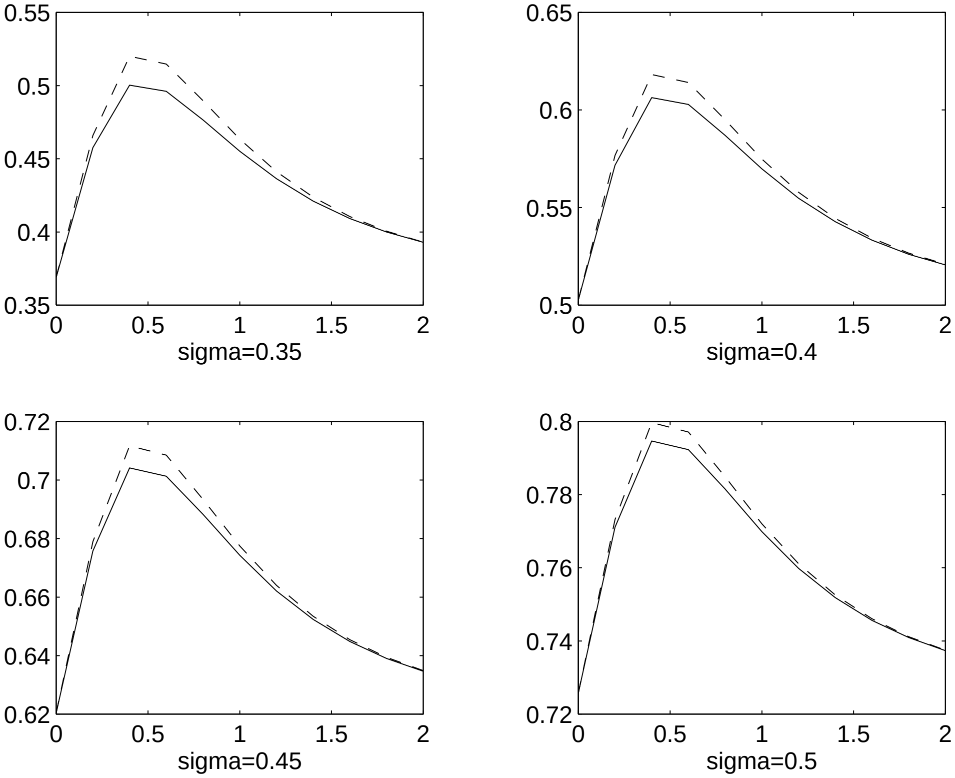

Figure 3: Numerical evaluation of the entropy (continuos line) and approximation using (9): parameter space = [0.5;5].

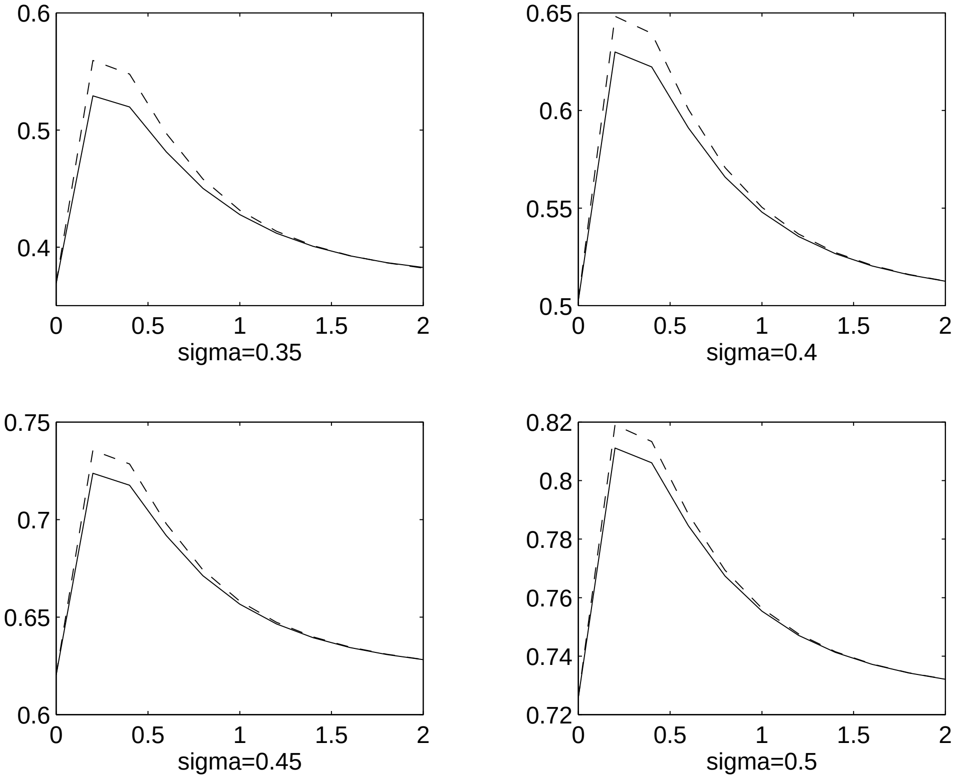

Figure 4: Numerical evaluation of the entropy (continuos line) and approximation using (9): parameter space = [0.5,10].

to justify an asymptotic approximation, thus asymptotically optimal design can be extremely ineffective.

5. Discussion

The MES principle allows us to find optimal design of experiments in a Bayesian framework with no need for asymptotic approximations of the design criterion. This result is particularly valuable when the experimenter is asked to design experiments based a small number of observations. The feasibility of the approach is limited to problems in which efficient numerical integrations techniques can be applied. For more complex problems a stochastic evaluation of the integrals can be investigated, and the fact that the MES principle allows us to avoid computations with conditional distributions can speed up computations of methods recently proposed, in which MCMC methods have been used to evaluate pre-posterior entropy (Müller & Parmigiani, 1995).

Acknowledgments

The first author acknowledges support from the Nuffield Foundation. The paper was presented at the ISI satellite meeting ISBA97, Turkey, and will appear in the Proceeding of the ASA 97: Section on Bayesian Statistics.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer-Verlag: New York. First Ed 1980.

Chaloner, K. (1993). A note on optimal Bayesian design for nonlinear problems. *J.Statist.Plan.Inf.*, *37*, 229–235.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, *10*, 273–304.

Cover, T. M., & Thomas, M. (1991). *Elements of Information Theory.* Wiley: New York.

Good, I. J. (1952). Rational decisions. *J.Roy.Statist.Soc.B*, 107–114.

Kitsos, C. P. (1995). On the support points of D-optimal non-linear experimental design for kinetics. In Kitsos, C. P., & Müller, W. G. (Eds.), *MODA4: Advances in model oriented data analysis, Sptses, Greece.* Physical Verlag: Heidelberg.

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

Müller, P., & Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J.Amer.Statist.Assoc.*, *90.*

Sebastiani, P., & Settimi, R. (1997). A note on d-optimal designs for a logistic regression models. *J.Statist.Plann.Inf.*, *59*, 359–368.

Sebastiani, P., & Settimi, R. (1998). First-order optimal designs for non-linear models. *J.Statist.Plann.Inf.*, To appear.

Sebastiani, P., & Wynn, H. P. (1997). Maximum entropy sampling and optimal Bayesian experimental design. *J.Roy.Statist.Soc.B*, To appear.

Sebastiani, P., & Wynn, H. P. (1998). Risk based optimal designs. *MODA5: Advances in model oriented data analysis.*, To appear.

Shewry, M. C., & Wynn, H. P. (1987). Maximum entropy sampling. *J.Appl.Statist.*, *14*, 165–170.

Sloan, I. H., & Joe, S. (1994). *Lattice Methods for Multiple Integration.* Clarendon Press: Oxford.