Biometrics 59, 1082–1088
December 2003

A Bayesian *A*-Optimal and Model Robust Design Criterion

Xiaojie Zhou,[1,*] Lawrence Joseph,[2,] David B. Wolfson,[3,***] and Patrick Bélisle,[4,****]**

[1]The Procter & Gamble Company, 8700 Mason-Montgomery Road, Mason, Ohio, U.S.A.
[2]Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada, and Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Quebec H3A 1A2, Canada
[3]Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke Street West, Montreal, Quebec H3A 2K6, Canada
[4]Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada
[*] *email:* zhou.x@pg.com
[**] *email:* Lawrence.Joseph@mcgill.ca
[***] *email:* david@math.mcgill.ca
[****] *email:* belisle@hirondeau.ri.mgh.mcgill.ca

Summary. Suppose that the true model underlying a set of data is one of a finite set of candidate models, and that parameter estimation for this model is of primary interest. With this goal, optimal design must depend on a loss function across all possible models. A common method that accounts for model uncertainty is to average the loss over all models; this is the basis of what is known as Läuter's criterion. We generalize Läuter's criterion and show that it can be placed in a Bayesian decision theoretic framework, by extending the definition of Bayesian *A*-optimality. We use this generalized *A*-optimality to find optimal design points in an environmental safety setting. In estimating the smallest detectable trace limit in a water contamination problem, we obtain optimal designs that are quite different from those suggested by standard *A*-optimality.

Key words: *A*-optimality; Bayesian optimal design; Decision theory; Model robustness.

1. Introduction

Optimal design can be used to sharpen the inferences that will be carried out following an experiment. There is an extensive literature on this subject when the underlying model is presumed known, but far less has been written on optimal design when there is uncertainty about the underlying model.

In this article we examine the question of robust optimal design under model uncertainty. When the model is unknown, the design criterion must work well regardless of which model is correct. Therefore, the optimal design should ensure that, on average, our estimator performs well over all models and, in addition, will not vary much amongst those models with high posterior probability. It is in this sense that the design criterion we propose is robust.

We begin by describing how optimal design might be employed in a particular environmental safety application. Frequently, it is desired to ascertain whether a given sample, for example, water, is contaminated in some way. Possible contaminants include bacteria and chemical substances that arise from environmental pollution. Often these contaminants occur in trace amounts and their presence must be detected by sophisticated means, such as chromatography and spectrometry (Kurtz, 1985).

Whatever the method used, the contaminant is very rarely observed directly, but rather readings are taken on some different scale. The assumption is made that there is a parametric relationship between the level of contamination and the measurements actually recorded. The curve of the hypothetical function relating the trace contamination-measurement pair is called the *calibration curve* of the particular measurement method. Calibration curves are constructed from designed experiments in which samples with different known contamination levels are analyzed, and the calibration curve fitted to the data.

An important characteristic of any method of trace residue analysis is the smallest detectable trace level that can be ascertained. Therefore, it is important to choose a design that in some sense will optimally estimate the detection limit. Several authors in Kurtz (1985) describe various frequentist methods for determining the detection limit L_D. Whether the approach proposed is frequentist or is Bayesian, as we shall propose, there are several features inherent to the problem. First, the selection of known trace amount levels for the experiment to determine the L_D is important (Currie, 1985; Mitchell, 1985). Second, a model must be specified relating the trace contamination amount to the measurement recorded. Although this is

commonly taken to be a simple linear model, it is recognized that model uncertainty can affect the results, and quadratic models have frequently been proposed as alternatives to linear models (Kurtz, Rosenberger, and Tamayo, 1985). Finally, note that the parameter L_D is common to all models, i.e., the interpretation of L_D remains the same from model to model.

The design problem here is to select experimental contamination values from the range of all possible values, in order to optimally estimate L_D. Since we are uncertain about whether the model is linear or quadratic, our design criterion must work well with either model. In analogy to frequentist terminology, we consider both "variance" and "bias," where with "variance," we refer to the performance of the selected design within each model, with "bias," refer to the performance of the design between models, i.e., how well the same design performs across a variety of plausible models. This viewpoint has also been taken by Box and Draper (1959), who show that, in the context of prediction, the "bias" can often be more important than the "variance." We shall show that by basing their design criteria on a weighted sum of losses of a certain form, other authors (Läuter, 1974, 1976; Cook and Nachtsheim, 1982) considered, implicitly, only "variance." We expand on these ideas in Section 2.

The outline of the article is as follows. Section 2 discusses the broader issue of optimal design in the presence of model uncertainty, beginning with a brief introduction to classical optimal design assuming the model is known. We review Läuter's criterion, which addresses a form of model uncertainty, and propose how it may be generalized. In Section 3, we show that the generalized Läuter's criterion may be formulated in a Bayesian decision theoretic context. We then define a generalized model robust Bayesian A-optimal design criterion. We return to the environmental water contamination problem to illustrate the difference between standard and generalized model robust A-optimality in Section 4, and concluding remarks are found in Section 5.

2. Optimal Design and Model Uncertainty

Most optimal design criteria assume that the model is known before the data are collected. In practice, however, we are often uncertain of the final model, even at the time of data analysis. From a frequentist viewpoint, a model usually refers to the likelihood function, but in the Bayesian paradigm, it includes not only the likelihood, but also the specification of a joint prior distribution over all unknown parameters. Model uncertainty may therefore arise from uncertainty about the prior distribution, the likelihood function, or possibly both. In the sequel, our use of the term "model" should be clear from the context, in that it will refer to the likelihood function alone when discussing frequentist inference, and to the likelihood/prior pair when discussing Bayesian methods. In either case, optimal designs derived from "model-known criteria," i.e., criteria that assume the correct model is fully known *a priori,* can depend strongly on the assumed model. In many situations, optimal designs concentrate observations at a small number of support points and can perform poorly if the assumed model is not correct. For example, in simple linear regression models when estimation of the slope is the goal, optimal designs tend to place design points at the extremes of the range of the independent variable. However, these designs can perform very poorly if the true model is in fact quadratic. Therefore, when the true model is unknown and there are parameters that retain their meaning from model to model, it is important that data be collected in such a way that the resulting parameter estimates are robust to model ambiguity.

Läuter (1974, 1976) allowed for model uncertainty in the choice of design by averaging design criterion functions over a finite set of possible models. This approach was adapted by Cook and Nachtsheim (1982) to optimal design for polynomial regression models when the degree of the polynomial is unknown, and by Dette (1993) for more general linear regression models. Optimal design for polynomial regression of unknown degree was more thoroughly investigated by Dette and Studden (1995) using as a criterion function the weighted p-mean of the relative efficiencies, which corresponds to Läuter's criterion when $p = 1$. They define relative efficiency of a design measure for a specified model in terms of the ratio of the determinant of the information matrix for the proposed design relative to that for the D-optimal design. The final criterion function of the design measure, ξ, is of the form $\{\sum_{i=1}^{l} w_i[\text{rel eff}_i(\xi)]^p\}^{1/p}$, where l is the upper bound of the degree of the polynomials considered plausible and w_i reflects a prior belief in the polynomial model of degree i.

A related but separate issue is that of optimal design for model identification. See, for example, Hill (1978). Our discussion here, however, focuses on robustness rather than model identification.

In the sequel, it is assumed that as we change from model to model, the parameters of main interest retain their interpretations. This is true in our water contamination application, for example, where the minimum detectable limit is the same regardless of the shape of the calibration curve.

When the correct model is assumed known *a priori,* classical optimal design theory may be summarized as follows: Assume that given a parameter $\theta \in \Theta$, the distribution of data $y \in \mathcal{Y}$ depends on the design variable $x \in \mathcal{X}$ and θ through a model M, usually defined through the likelihood function $f(y \mid \theta)$. The goal is to select a design x which will lead to optimal estimation of θ. Let $\hat{\theta}$ be an estimator of θ. A good design criterion ensures that a convex function $\phi(\Sigma(\hat{\theta}))$ of the covariance matrix $\Sigma(\hat{\theta})$ be as small as possible, on average, over the data y. Hence, an optimal design, x_{opt}, should minimize $\phi(\Sigma(\hat{\theta}))$. The most commonly used optimality criteria include D, A, and E-optimality, which correspond to ϕ being defined as log(det), tr and the maximum eigenvalue of $\Sigma(\hat{\theta})$, respectively, where det$(\cdot)$ is the determinant function and tr$(\cdot)$ is the trace of a matrix. Minimizing $\phi(\Sigma(\hat{\theta}))$ is desirable in D, A, and E-optimality, since small values of ϕ are associated with small values of the variance of the estimators. See Silvey (1980).

In standard linear regression problems, optimal designs can be found that are independent of the true value of the unknown parameter θ. In nonlinear problems, however, optimal designs usually depend on θ. Bayesian and frequentist methods handle these nuisance parameters differently. In the above frequentist setting, unknown values of θ are given fixed values that are hopefully close enough to the truth to provide good designs. The only integration that occurs, then, is in averaging the criterion function, $\phi(\Sigma(\hat{\theta}))$, over all possible data sets, according to the likelihood function with the assumed value

of θ. On the other hand, given a design x, Bayesian methods form the predictive distribution for the data y based on a prior distribution for θ, $f(y \mid x) = \int f(y \mid \theta, x) f(\theta)\, d\theta$, where $f(y \mid x)$ in the predictive distribution given design x, $f(y \mid \theta, x)$ is the likelihood function given design x, and $f(\theta)$ is the prior distribution for θ. Here, a design is typically chosen that minimizes the design criterion $\phi(\Sigma(y))$, where one averages over the distribution of the data y implied by $f(y \mid x)$, and $\Sigma(y)$ is the posterior covariance matrix of θ. Hence, there are two levels of integration, first over the value of θ, according to the posterior distribution of θ given each possible data set y, $f(\theta \mid y, x)$, and then over the unknown data y, according to the predictive distribution $f(y \mid x)$. Whether frequentist or Bayesian, standard optimal design criteria all assume that $f(y \mid \theta, x)$ and, in the case of Bayesian criteria, that $f(\theta)$ are known.

Läuter (1974, 1976) proposed the following criterion when the "true" unknown model is one among m candidate models, $\{M_1, M_2, \ldots, M_m\}$, for some finite m. We shall adapt Läuter's criterion to a Bayesian setting and then generalize our Bayesian version of Läuter's criterion.

Läuter's criterion: Let x be a given design, M be a given model (likelihood function). Let $a(M)$ be the action (estimator) taken under Model M. Let $\phi_{i,i}(x, a(M_i))$ denote the design criterion function whose arguments are as follows: the first subscript i, indicates that the predicted data are averaged over model M_i. The second subscript i indicates that the action $a = a(M_i)$ selected is also determined by the "presumed true" model M_i. Finally, the anticipated data are assumed to be collected under design x. Suppose further that the "true" model is, in fact, unknown and is one among m candidate models $\{M_1, M_2, \ldots, M_m\}$ for some finite m. Given weights w_i, an optimal design, x_{opt} is a design that optimizes (say, minimizes) the criterion function

$$\psi(x) = \sum_{i=1}^{m} w_i \phi_{i,i}(x, a(M_i)), \tag{1}$$

over all possible designs $x \in \mathcal{X}$.

Thus, the function $\psi(x)$ in (1) is composed of two averages. First, $\phi_{i,i}(x, a(M_i))$ is implicitly an average risk over the data predicted by model M_i when action $a(M_i)$ is taken. The two subscripts for ϕ are used to emphasize that the predicted data are generated under model M_i, which is also used to determine the action $a(M_i)$. The reason for this apparent pedantry will become clear in the generalization which follows. Second, $\psi(x)$ accounts for the uncertainty in the choice of model by providing a weighted average over the m models.

Now, even though (1) was originally proposed by Läuter as a frequentist criterion function, it may be easily interpreted as a Bayesian criterion function. The weight w_i may be interpreted as corresponding to the prior probability of model M_i, $i = 1, 2, \ldots, n$, and the model is assumed to include both a prior and a likelihood specification. We refer to this interpretation as Läuter's Bayesian criterion function, and assume henceforth that all discussion is within the Bayesian paradigm. *Once the data are collected,* however, it is natural to first compute parameter estimates predicated on each model and then account for model uncertainty by averaging these estimates with respect to the posterior model probabilities. Since, presumably, the data have in fact been generated by a unique model, in some instances few, if any, of the model specific estimates may be optimal.

This possibility arises because in the eventual analysis, the action taken may not be induced by the "true" model that generated the data. To guard against this eventuality, we should select a design that is robust in two senses:

(i) Actions $a(M_i)$ induced by data from models M_i should, on average, be "good" for estimating θ. The average here is a weighted average with weights w_i.

(ii) Estimates using actions $a(M_j)$, $j = 1, 2, \ldots, m$ should not differ much from one another, especially for those models that have the highest posterior probabilities of being selected (see Theorem 2 below). That is, in an average sense, we should not risk much by selecting a model M_j and basing our inferential procedure on it, even though our data were generated under model M_i, $i \neq j$, provided that M_i and M_j have relatively high posterior probabilities of being selected.

By choosing the criterion functions $\phi_{i,i}$ of Läuter's Bayesian criterion appropriately, we may account for robustness criterion (i). In order to include criterion (ii), however, we need an extended definition of the $\phi_{i,i}$'s.

Generalized Läuter's Bayesian Criterion: Let $\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ be the vector of prior probabilities corresponding to models $M_1, M_2, \ldots, M_m$. Let $\phi_{i,\underset{\sim}{M}} = \phi_i(x, \underset{\sim}{a}, \underset{\sim}{\pi})$ denote the risk function associated with using the procedures $a(M_j)$, $j = 1, 2, \ldots, m$, on data collected under M_i, where $\underset{\sim}{a} = (a_1(M_1), a_2(M_2), \ldots, a_m(M_m))$, and $\underset{\sim}{M} = 1, 2, \ldots, m$. The ith criterion function $\phi_{i,\underset{\sim}{M}}$ differs from $\phi_{i,i}$ in that instead of anticipating that the action will necessarily be based on the "true" model M_i, our eventual action may be based on any one of the m candidate models, whose precedence will depend on the prior probabilities $\underset{\sim}{\pi}$ through the obtained posterior probabilities.

DEFINITION: *The criterion function*

$$\psi_{\mathrm{GL}}(x) = \sum_{i=1}^{m} w_i \phi_{i,\underset{\sim}{M}}(x, \underset{\sim}{a}, \underset{\sim}{\pi})$$

is called the generalized Läuter's Bayesian criterion (GLBC) function. A criterion that depends on the optimization (say, minimization) of ψ_{GL} with respect to x is called a GLBC. In Section 3, *we move from the above generalities to show a precise link between the GLBC and a fully Bayesian decision theoretic approach.*

3. A Bayesian Decision Theoretic Formulation

Let $\theta \in \Theta \subset \mathcal{R}^p$ be an unknown parameter whose meaning is independent of the choice of model M, as discussed in Section 2. Let $a(M) \in \Theta$ denote an estimator of θ. In practice, $a(M)$ evaluated at (x, y), namely, $a(M, x, y)$, will depend on the observed data y whose prior predictive distribution will depend on the prior distribution and on the design x.

Define $\overline{a} = \sum_{j=1}^{m} a(M_j) P(M = M_j \mid y)$, where $a(M_j)$ is the estimator of θ selected under model M_j. Let $L = L_{M_j}(\theta, x, y, \overline{a}) \in \mathcal{R}$ be the loss at $\theta \in \Theta$ associated with the design x, the observed data y collected under M_j, and the average action $\overline{a}$. Note that we allow the alternative actions $a(M_j)$,

mentioned in Section 2 to enter the loss function through the weighted average, $\overline{a}$. Given a data set y, define

$$E_{M,\theta|y}(L_{\theta,x,y,\overline{a}}) = \sum_{j=1}^{m} \int_{\Theta} L_{M_j}(\theta, x, y, \overline{a}) P(M = M_j \,|\, y) \times f(\theta \,|\, M = M_j, y)\, d\theta.$$

Define the risk, $R(x, \overline{a})$, associated with design x as the average of the above expected loss over the predictive distribution of y, given by

$$R(x, \overline{a}) = \int_{\mathcal{Y}} E_{M,\theta\,|\,y}(L_M(\theta, x, y, \overline{a})) f(y \,|\, x)\, dy, \qquad (2)$$

where

$$f(y \,|\, x) = \sum_{j=1}^{m} \int_{\Theta} f(y \,|\, x, \theta, M) f(\theta \,|\, M_i)\, d\theta \times \pi_j$$

is the prior predictive distribution of y for design x, averaged over all models, M_j, $j = 1, 2, \ldots, m$, and where $\pi_j = P(M = M_j)$. Theorem 1 below states that the GLBC function, $\psi_{\mathrm{GL}}(x)$, can be expressed as the risk, $R(x, \overline{a})$, as defined above, so that the GLBC can be interpreted as an expected loss in a Bayesian decision theoretic setup.

THEOREM 1: *Given M_i, for $i = 1, 2, \ldots, m$, let*

$$\phi_{i,\underset{\sim}{M}} = \phi_i(x, \underset{\sim}{a}, \underset{\sim}{\pi}) = \int_{\mathcal{Y}} \left\{ \int_{\Theta} L_{M_i}(\theta, x, y, \overline{a}) f(\theta \,|\, M_i, y)\, d\theta \right\} f(y \,|\, M_i, x)\, dy.$$

Then $R(x, \overline{a})$, as defined by (2) is expressible as

$$R(x, \overline{a}) = \sum_{i=1}^{m} w_i \phi_i(x, \overline{a}, \underset{\sim}{\pi}),$$

for weights w_i, $i = 1, 2, \ldots, m$, where $w_i = \pi_i$.

Proof. By definition,

$$\begin{aligned} R(x, \overline{a}) &= \int_{\mathcal{Y}} E_{M,\theta\,|\,y}\left(L_M(\theta, x, y, \overline{a})\right) f(y \,|\, x)\, dy \\ &= \int_{\mathcal{Y}} \sum_{j=1}^{m} \int_{\Theta} L_{M_j}(\theta, x, y, \overline{a}) P(M = M_j \,|\, y) \\ &\quad \times f(\theta \,|\, M = M_j, y)\, d\theta f(y \,|\, x)\, dy \\ &= \sum_{i=i}^{m} \int_{\mathcal{Y}} \int_{\Theta} L_{M_i}(\theta, x, y, \overline{a}) f(\theta \,|\, M_i, y) f(y \,|\, M_i, x)\, d\theta\, dy \times \pi_i \\ &= \sum_{i=i}^{m} \pi_i \int_{\mathcal{Y}} \int_{\Theta} L_{M_i}(\theta, x, y, \overline{a}) f(\theta \,|\, M_i, y) f(y \,|\, M_i, x)\, d\theta\, dy \\ &= \sum_{i=i}^{m} w_i \phi_i(x, \overline{a}, \underset{\sim}{\pi}). \end{aligned}$$

COROLLARY: *Suppose that θ is scalar valued and that*

$$L_{M_i}(\theta, x, y, \overline{a}) = \lambda E_{\theta\,|\,M_i}\left\{\theta - E_{\theta\,|\,M_i}(\theta \,|\, x, y, M_i)\right\}^2 + (1-\lambda)\left\{E_{\theta\,|\,M_i}(\theta \,|\, x, y, M_i) - E(\theta \,|\, x, y)\right\}^2,$$

where

$$E(\theta \,|\, x, y) = \sum_{j=1}^{m} E(\theta \,|\, x, y, M_j) \times P(M = M_j \,|\, x, y),$$

and where $0 \le \lambda \le 1$. Then the GLBC function may be expressed as

$$\int_{\mathcal{Y}} [\lambda E_M\{\mathrm{var}_{\theta|M}(\theta \,|\, x, y, M)\} + \{(1-\lambda)\mathrm{var}_M(E_{\theta|M}(\theta \,|\, x, y, M))]\} f(y \,|\, x)\, dy. \qquad (3)$$

More generally, if $\theta \in \mathcal{R}^p$ is vector valued, the integrand of (3) becomes

$$\mathrm{tr}[\lambda E_M(\mathrm{covar}_{\theta\,|\,M}(\theta \,|\, x, y, M)) + (1-\lambda)(\mathrm{covar}_M E_{\theta|M}(\theta \,|\, x, y, M))], \qquad (4)$$

where $\mathrm{tr}(\cdot)$ is the trace function and $\mathrm{covar}(\cdot)$ denotes the covariance matrix function.

With $\lambda = 1$, the expression (4) reduces to the expected sum of the posterior variances of the components of θ, corresponding to the weighted sum of p Bayesian A-optimal design criterion functions. We propose, therefore, a generalized robust Bayesian A-optimal design criterion:

Generalized Robust Bayesian A-Optimality: Let $M \in \mathcal{M}$ be a class of possible models for the pair (θ, y). A design x_{opt} is said to be a generalized robust Bayesian A-optimal design if x_{opt} minimizes

$$\int_{\mathcal{Y}} \mathrm{tr}[\lambda E_M(\mathrm{covar}_{\theta\,|\,M}(\theta \,|\, x, y, M)) + (1-\lambda)(\mathrm{covar}_M E_{\theta|M}(\theta|x, y, M))] f(y \,|\, x)\, dy.$$

A design x_{opt} that is optimal in the sense of generalized robust Bayesian A-optimality is robust against model uncertainty in two senses:

(i) On average, the posterior variance of θ is small when the model that generates the data and that actually is used to derive the posterior distribution coincide, and
(ii) on average, over the different models, the posterior parameter means have small variance.

Consequence (ii) implies that the posterior means used to estimate θ should not differ much from one another even when their computation is based on misspecified models for (y, θ). This is especially important among models with high posterior probabilities. This is made more precise in Theorem 2 below.

THEOREM 2: *Let $\tilde{w}_i$ be the posterior probability that M_i is the true model for the data y collected at design x, $i = 1, \ldots, m$. Then*

$$\mathrm{var}_M(E_{\theta|M}(\theta \,|\, x, y, M)) = \sum_{i<j} \tilde{w}_i \tilde{w}_j \left\{E_{\theta|M_i}(\theta \,|\, x, y, M_i) - E_{\theta|M_j}(\theta \,|\, x, y, M_j)\right\}^2.$$

Proof. To simplify notation, we use E_i to denote $E_{\theta|M_i}(\theta \mid x, y, M_i)$. Then

$$\begin{aligned}&\mathrm{var}_M(E(\theta \mid y, x, M))\\ &= \sum_{i\neq j} \tilde{w}_i\tilde{w}_j E_i(E_i - E_j)\\ &= \sum_{i\neq j} \tilde{w}_i\tilde{w}_j (E_i - E_j)^2 + \sum_{i\neq j} \tilde{w}_i\tilde{w}_j E_j(E_i - E_j).\end{aligned}$$

The last term in above expression can be written as

$$\sum_{i\neq j} \tilde{w}_i\tilde{w}_j E_j(E_i - E_j) = -\sum_{i\neq j} \tilde{w}_i\tilde{w}_j E_i(E_i - E_j).$$

Hence,

$$\begin{aligned}\mathrm{var}_M(E_{\theta|M}(\theta \mid x, y, M) &= \frac{1}{2}\sum_{i\neq j} \tilde{w}_i\tilde{w}_j (E_i - E_j)^2\\ &= \sum_{i<j} \tilde{w}_i\tilde{w}_j (E_i - E_j)^2.\end{aligned}$$

We may interpret Theorem 2 as follows: The expression $\sum_{i<j} \tilde{w}_i\tilde{w}_j(E_i - E_j)^2 = (1/2)\sum_{i,j} \tilde{w}_i\tilde{w}_j(E_i - E_j)^2$ may be written as $\sum_{i=1}^m \tilde{w}_i\overline{d}_i$ where $\overline{d}_i = \sum_{j=1}^m \tilde{w}_j(E_i - E_j)^2$ is the weighted sum of the distances of the E_j's from E_i. The weights are the posterior probabilities of the models M_j. Hence $\overline{d}_i$ describes a weighted average distance of the estimates E_j from E_i. The expression $\sum_{i=1}^m \tilde{w}_i\overline{d}_i$ is a weighted average of these distances. If $\tilde{w}_{(1)} < \tilde{w}_{(2)} < \cdots < \tilde{w}_{(m)}$, and $\overline{d}_{(1)} < \overline{d}_{(2)} < \cdots < \overline{d}_{(m)}$, are the ordered $\tilde{w}_i$'s and $\overline{d}_i$'s, respectively, then it can be shown (Hardy, Littlewood, and Polya, 1997) that $\sum_{i=1}^m \tilde{w}_i\overline{d}_i$ would be minimized over all possible permutations of the $\tilde{w}$ and $\overline{d}$ vectors by choosing the permutations that yield the products $\tilde{w}_{(i)}\overline{d}_{(m-i+1)}, i = 1, 2, \ldots, m$. That is, the minimum value of the sum would be $\sum_{i=1}^m \tilde{w}_{(i)}\overline{d}_{(m-i+1)}$. We cannot, of course, average our pairings in this optimal way, as for each $\tilde{w}_{(i)}$, there is a $\overline{d}_{(j_i)}$ that accompanies it. The ideal, however, is to choose the design x in such a way so that the sum is a minimum. In order for $\overline{d}_i$ to be small, $(E_i - E_j)^2$ should be small for j's such that $\tilde{w}_j$ is large. Alternatively, those estimates E_i which are furthest on average from the other estimates, E_j, should correspond to models with low posterior probability. A good design ensures that models with high posterior probability produce estimates that are on average close to those arising from other models with high posterior probability.

The Choice of λ*:* The choice of the weighting parameter λ should be determined in accordance with the relative emphasis to be placed on the robustness considerations (i) and (ii). By choosing $\lambda = 1/2$ in equation (3), we see that minimizing $R(x)$ is equivalent to minimizing $2R(x) = E_{Y|x}\{\mathrm{var}(\theta \mid x, Y)\}$. When $\lambda = 1$, the risk is simply $\int_y E_M(\mathrm{var}_{\theta|M}(\theta \mid x, y, M))f(y \mid x)\, dy$.

4. Optimal Design for Estimating the Detection Limit in a Water Contamination Experiment

While Mitchell (1985) does not specifically address optimal design, here we find the optimal design for estimating the detection limit, L_D, in the water contamination experiment that was described in the introduction. We present a Bayesian approach to the estimation of L_D, followed by a discussion of the optimal design.

Recall that the calibration curve can be either linear or quadratic. Regardless of the form of the curve, it is first necessary to find the smallest value on the (transformed) calibration scale that can be reasonably anticipated to be consistent with a nonzero value of contamination. We denote this value by y_0, and assume that it is known from previous experiments. It can be found, for example, by taking several measurements of the calibrated value, y, each corresponding to a value at $x = 0$, i.e., no contamination, and taking the upper $100(1 - \alpha)$th percentile of the posterior predictive distribution for y from these data.

First assume the model is linear, that is, assume

$$Y_i = a + bx_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where a and b are constants and the errors ϵ_i are independently and identically distributed $N(0, \sigma^2)$ random variables. The design point x_i corresponds to the selected trace contamination amount, and Y_i represents the observed measurement corresponding to x_i for the ith out of a total of n observation pairs.

Similarly, we can define a quadratic model,

$$Y_i = a + bx_i + cx_i^2 + \epsilon_i, \quad i = 1, 2, \ldots, n\,,$$

where a, b, and c are constants and ϵ_i are again independently and identically distributed $N(0, \sigma^2)$ random variables. The goal is to estimate L_D, defined as $(y_0 - a)/b$ for the linear model, and as the smallest positive solution to the quadratic equation $y_0 = a + b \times L_D + c \times L_D^2$ for the quadratic model.

Following Mitchell (1985), suppose that the design space is $\mathcal{X} = \{0, 10, 20, 30, 40, 50, 60\}$, and we wish to choose a total of $n = 8$ design points within this set, with repeated choices allowed. This implies a total of 3303 unique designs, and the question is which of these provides optimal estimation of L_D. We use generalized A-optimality as our design criterion, and show that the optimal design varies as λ varies across the range from 0 to 1.

Roughly following the example on page 123 of Mitchell (1985), we used the following prior distributions. The intercept for the linear model was taken to be $a \sim N(2, 25)$, with slope $b \sim N(0.3, 0.09)$. Further, we took $\sigma^2 \sim \mathrm{IG}(1.25, 0.6)$, where the latter denotes an inverse gamma distribution with mean 2.4, with parameters chosen such that the regression standard deviation has a 99% credible set that runs approximately from 0 to 6. For the quadratic model, we used $a \sim N(3, 25)$, $b \sim N(0.7, 0.09)$, $c \sim N(-0.005, 0.000004)$, and $\sigma^2 \sim \mathrm{IG}(1.25, 0.6)$. We assumed that all parameters were *a priori* independent, allowing the data to provide the appropriate correlations between the parameters within each model *a posteriori*. In rare cases, these prior distributions gave rise to lines or curves where no real solution within the range from 0 to 60 could be found (for example, due to negative slopes or decreasing relationship between x and y). These cases were excluded from all analyses described below.

For each model, we calculated the risk as given by (4) where, in our context, θ was replaced by L_D. We calculated this risk for $\lambda = 0$, 0.25, 0.5, 0.75 and 1. As closed

form formulae for the covariances required by (4) were not available, we used the Gibbs sampler, as programmed in WinBUGS (Spiegelhalter et al., 1997), which provided estimates for $\psi_{\mathrm{GL}}(x)$ for all 3303 possible designs, for each λ value in the set $\lambda = (0, 0.25, 0.5, 0.75, 1)$. While an exhaustive search was carried out among all designs, as each risk estimate is only approximate, it does not necessarily follow that the designated optimal design corresponding to each value of λ is the true optimal design, due to approximation error in the WinBUGS simulations. It was not feasible to run sufficient numbers of iterations across all 3303 designs to obtain exact optimal designs. We therefore followed the following strategy to obtain approximate optimal designs for each value of λ. We first ran 3000 iterations for all 3303 designs, and found the top 50 designs for each of the five values of λ. For these 250 designs (in fact slightly fewer designs, since there were some designs which were in the top 50 for more than one value of λ), we reran the analyses with 5000 iterations; we next took the top 10 designs for each value of λ, which resulted in 31 unique designs. Finally, we ran 15,000 iterations for these designs, and selected the top 3 from each value of λ to report. While there is no guarantee that the absolute optimal design is among those included in our final top 31 designs, as we discuss below, the strong trends in the results allows for useful generalizations to be stated about which types of designs will perform well for the various values of λ.

The results are presented in Table 1, which shows that optimal designs from the generalized Bayesian A-optimality criterion with $\lambda = 0.5$ leads to designs that are quite different from those induced by standard Bayesian A-optimality

Table 1

Risk results for the minimal detection limit problem described in Section 4. For each design, we present the average risk across the five λ values, including 0, 0.25, 0.5, 0.75, *and* 1. *The best three designs for each value of λ are shown.*

		Design points						
Rank	Risk	0	10	20	30	40	50	60
				$\lambda = 0$				
1	0.01838	3	0	0	2	1	0	2
2	0.02062	3	0	0	2	0	2	1
3	0.02313	3	1	0	1	2	0	1
				$\lambda = 0.25$				
1	0.19167	5	0	0	1	1	0	1
2	0.21307	5	0	0	1	0	1	1
3	0.23684	4	1	0	2	0	0	1
				$\lambda = 0.5$				
1	0.35869	5	0	0	1	1	0	1
2	0.38452	5	0	0	1	0	1	1
3	0.40075	5	0	0	0	1	1	1
				$\lambda = 0.75$				
1	0.45855	6	0	0	0	0	0	2
2	0.52571	5	0	0	1	1	0	1
3	0.53876	5	0	0	0	0	0	3
				$\lambda = 1$				
1	0.44657	6	0	0	0	0	0	2
2	0.53896	5	0	0	0	0	0	3
3	0.63259	7	0	0	0	0	0	1

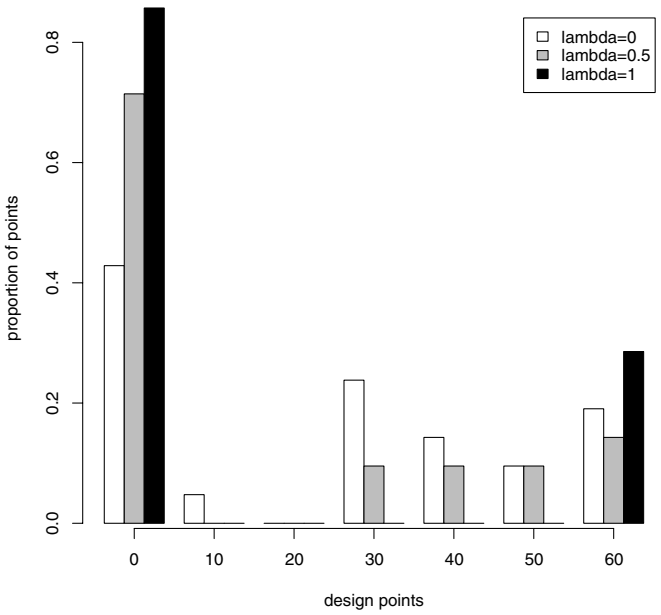

Figure 1. Graphical representation of the proportion of design points at each of the 7 possible positions for each of the three values of $\lambda = 0$, 0.5, and 1, for the detection problem described in Section 4.

($\lambda = 1$). The optimal designs under $\lambda = 1$ tend to place all design points at the extremes of the design space, at $x = 0$ and $x = 60$, with the top three designs including no other points. This is similar to results from standard optimal design in linear regression, where points at the extremes are often optimal. In contrast, for $\lambda = 0.5$, several non-boundary points are also included, with even more non-boundary points included for $\lambda = 0$. This is as expected, since if one needs to ensure reasonable estimation across both linear and quadratic models, one needs to not only estimate the slope of a linear model accurately, as dictated by points near the ends of the design space (see Silvey, 1980), but also the curvature of the quadratic model, so that points away from the boundary are also needed. Designs intermediate to those just described are seen when $\lambda = 0.25$ and $\lambda = 0.75$. Figure 1 summarizes the results by presenting the average proportion of points at each possible location for λ values of 0, 0.5, and 1. Clearly, more nonboundary points are included when robustness is considered.

Overall, these designs are intuitively sensible, with most points placed near $x = 0$, which is close to the area of greatest interest, i.e., close to L_D. All designs also include one or more points near $x = 60$, required to ensure estimation of the slope or curve of the calibration function. The degree to which inner points are also included depends on the degree to which robustness is considered.

5. Concluding Remarks

We propose a Bayesian generalization of Läuter's criterion for robust optimal design when there is model uncertainty. This criterion is then placed in a Bayesian decision theoretic framework and is shown to suggest a generalized robust Bayesian A-optimal design criterion. Using generalized

Bayesian *A*-optimality to find the optimal design in a water contamination problem, we obtain optimal designs that are quite different from those resulting from standard Bayesian *A*-optimality. Special cases subsumed by our discussion are problems containing a nuisance parameter whose set of possible values is finite, as well as finite mixture problems. Extensions of these to their respective infinite valued parameter space counterparts remain open questions.

As Key, Pericchi, and Smith (1999) point out, only rarely does one believe that the list of m models under consideration includes the model which exactly describes the real world mechanism that generated the data. Nevertheless, one can consider a set of models, one or more of which can be close enough to this ideal to provide useful answers to the research question. Generalized Bayesian *A*-optimality can be applied to either of these perspectives, so that one does not necessarily need to assume that the "correct" model is among those being considered. While we assume that the parameters retain their meanings from model to model, this is a sufficient, but not necessary, condition. Rather, it is sufficient that the loss functions from each model are comparable (that is, are on similar scales) to each other.

Our work is related to that of Raftery, Madigan, and Hoeting (1997) on Bayesian model averaging for regression problems, where predictions are based on the weighted average prediction over all plausible models, the weights being the posterior probabilities of these models. Our criterion can therefore be useful in designing studies in which Bayesian model averaging will be used.

Résumé

Supposons que le véritable modèle sous-tendant un jeu de données fasse partie d'un ensemble fini de modèles candidats, et que notre intérêt se porte en premier lieu sur l'estimation des paramètres dudit modèle. En fonction de cet objectif, un plan optimal doit dépendre d'une fonction de perte définie pour tous les modèles possibles. Une méthode classique, qui tient compte de l'incertitude sur le modèle, consiste à calculer la perte moyenne sur l'ensemble de tous les modèles envisagés – cette méthode constitue le fondement d'un critère connu sous le nom de ≪critère de Läuter≫. Nous généralisons ce critère de Läuter, et démontrons qu'il est possible, en élargissant la définition de l'A-optimalité bayésienne, de le resituer dans un cadre théorique de décision bayésienne. A l'aide de cette A-optimalité généralisée, nous pouvons ainsi définir les points expérimentaux optimaux d'un problème de sécurité environnementale, où, pour estimer le seuil de détection d'une contamination de l'eau, nous obtenons des plans optimaux très différents des plans suggérés par l'A-optimalité bayésienne standard.

References

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association* **54,** 622–654.

Cook, R. D. and Nachtsheim, C. J. (1982). Model-robust, linear optimal designs. *Technometrics* **24,** 49–54.

Currie, L. (1985). The many dimensions of detection in chemical analysis with special emphasis on the one-dimensional calibration curve. In *Trace Residue Analysis*, Kurtz, D. (ed), 49–81. Washington, D.C.: American Chemical Society.

Dette, H. (1993). Bayesian D-optimal and model robust designs in linear regression models. *Statistics* **25,** 27–46.

Dette, H. and Studden, W. J. (1995). Optimal designs for polynomial regression when the degree is not known. *Statistica Sinica* **5,** 459–473.

Hardy, G., Littlewood, J. E., and Polya, G., (1997). *Inequalities,* 2nd edition. Cambridge: Cambridge University Press.

Hill, P. D. H. (1978). A review of experimental design procedures for regression model discrimination. *Technometrics* **9,** 57–71.

Key, J., Pericchi, L., and Smith A. F. M. (1999). Bayesian model choice: What and why? (with discussion). In *Bayesian Statistics 5,* J. Bernardo et al. (eds), 343–370. Oxford: University Press.

Kurtz, D. (1985). *Trace Residue Analysis*. Washington, D.C.: American Chemical Society.

Kurtz, D., Rosenberger, J., and Tamayo, G. (1985). The linear calibration graph and its confidence bands from regression on transformed data. In *Trace Residue Analysis,* D. Kurtz (ed), 133–165. Washington, D.C.: American Chemical Society.

Läuter, E. (1974). Experimental design in a class of models. *Mathematische Operationsforschung und Statistik* **5,** 379–396.

Läuter, E. (1976). Optimal multipurpose designs for regression models. *Mathematische Operationsforschung und Statistik* **7,** 51–68.

Mitchell, D. (1985). Calibration-based-curve analysis: Use of multiple curve and weighted least squares procedures with confidence band statistics. In *Trace Residue Analysis,* D. Kurtz (ed), 115–131. Washington, D.C.: American Chemical Society.

Raftery, A., Madigan, D., and Hoeting J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92,** 179–191.

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

Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. (1997). *BUGS 0.6—Bayesian Inference Using Gibbs Sampling*. Cambridge, U.K.: MRC Biostatistics Unit.

Received January 2001. *Revised April* 2003.
Accepted April 2003.