Design of Follow-Up Experiments for Improving Model Discrimination and Parameter Estimation

Szu Hui Ng[1] • Stephen E. Chick[2]

National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260.
Technology Management Area, INSEAD, 77305 Fontainebleau CEDEX, France.
isensh@nus.edu.sg • stephen.chick@insead.edu

One goal of experimentation is to identify which design parameters most significantly influence the mean performance of a system. Another goal is to obtain good parameter estimates for a response model that quantifies how the mean performance depends on influential parameters. Most experimental design techniques focus on one goal at a time. This paper proposes a new entropy-based design criterion for follow-up experiments that jointly identifies the important parameters and reduces the variance of parameter estimates. We simplify computations for the normal linear model by identifying an approximation that leads to a closed form solution. The criterion is applied to an example from the experimental design literature, to a known model and to a critical care facility simulation experiment.
(*Design of Experiments; Model discrimination; Parameter estimation; Entropy; Simulation*)

1. Introduction

A common purpose of many experiments is to obtain an adequate mathematical model of the underlying system, including the functional form, and precise estimates of the model's parameters. Response models that describe the relationship between inputs to a system and the output can be useful for design decisions, and much focus has gone into selecting inputs in a way that improves the estimate of the response model [9, 29, 30, 40]. Response models can be used in iterative processes to identify design parameters (e.g., number of servers, production line speeds) that optimize some expected reward criterion (e.g., mean monthly revenue, average output), or to provide intuition about how input factors influence aggregate system behavior. In simulation, response models can relate the parameters of stochastic models (e.g., demand arrival rates, infection transmission parameters) to system performance [2, 13, 24, 25, 32, 35].

[1]Corresponding author.
[2]The authors acknowledge the financial support of the National Institutes of Health (grant R01 AI45168-01A1).

A number of design criteria are available to select design factors (or inputs) for experiments. Several authors [6, 16, 26] use the expected gain in Shannon information (or decrease in entropy) as an optimal design criterion to select values for the experiment's design factors. Bernardo [6] and Smith and Verdinelli [38] adopted this approach and looked at how to plan experiments to ensure precise estimates of the model's parameters. However, many experiments aim to identify which factors most influence the system response. Identifying the subset of most important parameters can be phrased as a model selection problem [34]. Box and Hill [10] used Shannon information to develop the MD design criterion for discriminating among multiple candidate models.

Hill [21] stressed the importance of experimental design for the joint objective of model discrimination and parameter inference in his review of design procedures. But most design criteria focus either on identifying important parameters or improving estimates of response parameters, but not both. Exceptions are Hill et al. [22], whose joint criterion requires certain model parameters to be estimated or known, and Borth [8], whose entropy-based criterion can be challenging to compute.

This paper describes a new joint criterion for experimental design that selects designs to simultaneously identify important factors and to reduce the variance of the response model parameter estimates. The new criterion is shown to simplify to a closed form for the standard linear regression model with normal observation errors, and is computationally more efficient than Borth's criterion. Our criterion does not require initial estimates of the model parameters and incorporates prior information and data from preliminary experiments. It is flexible for use in either starting or follow-up experiments, particularly if results remain inconclusive about which factors most influence the system response, and when the parameters are still poorly understood after an initial response surface experiment has been completed. We consider designs with a given number n of observations, and do not describe how to balance initial runs with follow-up runs.

Section 2 describes the mathematical formulation for the design space and response models. It also describes a Bayesian formulation to quantify input model and parameter uncertainty, as well as the new entropy based design criterion. Three numerical experiments in Section 3 show that optimal designs depend heavily on the criterion selected, and highlight the benefits and tradeoffs of the new joint criterion over individual model discrimination and parameter estimation criteria, as well as existing joint criteria. The examples stress the need for balancing the two types of entropy measures of the joint criterion. For the examples considered, we also find that the weights in our criterion are robust to misspecification. The new criterion does well at both identifying important factors and reducing parameter uncertainty, and is computationally more efficient than

Borth's joint criterion.

2. Formalism

The design criterion is applied to a finite, but perhaps large, set of potential factors in a finite number n of runs, where n is selected by the experimenter. The design space and class of regression models is described before the new entropy-based design criterion and computational issues.

2.1 Design Space and Regression Models

Experiments often involve several factors. Here we consider representing the performance of the systems by the usual linear model. We consider a finite number q of real-valued factor inputs, $x_1, \ldots, x_q$, each of which may be chosen to take on a finite set of different values. These factors can be combined algebraically to generate a finite number, p, of predictors, $y_1, \ldots, y_p$, each of which is some function of the input factors.

We follow the formulation of Raftery et al. [34] to identify the most important of the p predictors. That is, we presume the existence of $s = 2^p$ candidate response models in a model space $\mathbf{M}$ that are linear in some subset of the predictors. Assuming the ℓ-th candidate response model, the output z_i of the i-th run is presumed to be of the form

$$z_i = \beta_0 + \beta_1 y_{i,(1)} + \beta_2 y_{i,(2)} + \cdots + \beta_t y_{i,(t)} + \zeta_i, \tag{1}$$

where $y_{i,(1)}, \ldots, y_{i,(t)}$ are the t predictors present in the ℓ-th model, the values of the β_j may depend upon ℓ, and ζ_i is a zero mean noise term. The selection of a candidate response model identifies the important predictors, relative to the size of the noise in the response. See also George [20].

Let $\mathbf{D}$ be the (finite) design space of all possible legal combinations of the inputs for each of the n runs. A design $\mathbf{x} \in \mathbf{D}$ can be represented as an $n \times q$ matrix whose i-th row contains the values of the factors for the i-th run. If model M_ℓ has t predictors, the design matrix can be converted to an $n \times (t+1)$ predictor matrix $\mathbf{y}_\ell = \mathbf{y}_\ell(\mathbf{x})$ whose rows contain the values of predictors for each run, and the first column corresponds to the intercept. Let $\mathbf{z}$ be the column vector of n outputs.

2.2 Entropy-Based Formulation

The problem is to choose a design $\mathbf{x}$ that in some sense is effective at identifying the most important predictors (i.e., selects the most appropriate model in $\mathbf{M}$), and estimate regression parameters.

We assess uncertainty about response model selection and parameter estimation with probability distributions. The design that most improves an entropy-based criterion is then selected.

2.2.1 Uncertainty Assessment

One Bayesian approach to quantify the joint uncertainty about model form and parameter values is to assign a prior distribution to each of the models $M_\ell \in \mathbf{M}$, then assign a conditional probability distribution for the parameter vector $\boldsymbol{\beta}_\ell$, given M_ℓ. The identity of the best response model and parameter is then inferred by Bayes' rule, using the prior distributions and the probability distribution of the output, given the model and input parameters. This is the approach taken by [14, 28, 34].

We make a standard assumption of jointly independent, normally distributed errors, $\zeta_\ell \sim \text{Normal}\,(0, \sigma^2)$, so if model M_ℓ is the model, $\boldsymbol{\beta}_\ell$ is the parameter, $\mathbf{x}$ is the design with predictor matrix $\mathbf{y}_\ell = \mathbf{y}_\ell(\mathbf{x})$, then the output $\mathbf{Z}$ has an multivariate normal distribution,

$$p(\mathbf{Z} \mid M_\ell, \boldsymbol{\beta}_\ell, \sigma^2, \mathbf{x}) \sim \text{Normal}\left(\mathbf{y}_\ell \boldsymbol{\beta}_\ell, \sigma^2 I_n\right),$$

where I_n is the identity matrix. For prior distributions, we presume a conjugate prior distribution [5] for the unknown $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2)$, conditional on the ℓ-th model M_ℓ,

$$\begin{aligned} \pi(\boldsymbol{\beta}_\ell \mid M_\ell, \sigma^2) &\sim \text{Normal}\left(\boldsymbol{\beta}_\ell \mid \boldsymbol{\mu}_\ell, \sigma^2 \mathbf{V}_\ell\right) \\ \pi(\sigma^2 \mid M_\ell) &\sim \text{InvertedGamma}\left(\sigma^2 \mid \frac{\nu}{2}, \frac{\nu\lambda}{2}\right), \end{aligned} \tag{2}$$

where the conditional prior mean vector $\boldsymbol{\mu}_\ell$ and covariance matrix $\sigma^2 \mathbf{V}_\ell$ for $\boldsymbol{\beta}$ may depend on the model M_ℓ. The parameters ν and λ are selected by the modeler. The InvertedGamma $(x \mid \alpha, \beta)$ distribution has pdf $x^{-(\alpha+1)} e^{-\beta/x} \beta^\alpha / \Gamma(\alpha)$ and mean $\beta/(\alpha - 1)$. Raftery et al. [34] suggest values of $\boldsymbol{\mu}_\ell$, $\mathbf{V}_\ell$, ν and λ that minimize the influence of the priors in numerical experiments.

The distributions in Eq. (2) can either be based on prior information alone, or can include information gained during initial stages of experimentation. Data $\mathbf{z}_0$ from an initial stage of n_0 observations with predictor matrix $\mathbf{y}_0$ is straightforward to incorporate because of the conjugate form [5]. Replace the mean $\boldsymbol{\mu}_\ell$ with $\boldsymbol{\mu}'_\ell = \left(\mathbf{V}_\ell^{-1} + \mathbf{y}_0^{\mathrm{T}} \mathbf{y}_0\right)^{-1} \left(\mathbf{V}_\ell^{-1} \boldsymbol{\mu}_\ell + \mathbf{y}_0^{\mathrm{T}} \mathbf{z}_0\right)$; replace $\mathbf{V}_\ell$ with $\left(\mathbf{V}_\ell^{-1} + \mathbf{y}_0^{\mathrm{T}} \mathbf{y}_0\right)^{-1}$; replace $\nu/2$ with $(\nu + n_0)/2$; and replace $\nu\lambda/2$ with $(\nu\lambda + (\mathbf{z}_0 - \mathbf{y}_0 \boldsymbol{\mu}'_\ell)^{\mathrm{T}} \mathbf{z}_0 + (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}'_\ell)^{\mathrm{T}} \mathbf{V}_\ell^{-1} \boldsymbol{\mu}_\ell)/2$.

Choices used here for the prior distribution of M_ℓ include the discrete uniform ($p(M_\ell) = 1/s$) and the independence prior $p(M_i) = \omega^{t_i}(1 - \omega)^{p - t_i}$, where t_i is the number of predictors in model i, ($i = 1, \ldots, s$), and ω is the prior probability that a predictor is active. Raftery et al. [34] provide closed form formulas to update the probabilities $p(M_\ell \mid \mathbf{z}_0, \mathbf{y}_0)$.

In the rest of the paper, the prior distribution for the follow up stage is based on a prior distribution from Eq. (2) in combination with data from an initial stage. The optimal balance of the amount of initial stage data versus the amount of follow-up data is beyond the scope of this paper.

2.2.2 Modelling Remarks

Considering posterior probabilities is a useful way to assess the relative merits of the models [20]. Selecting models according to $p(M_\ell|\mathbf{Z})$ is consistent in that if one of the entertained models is actually the true model, then it will select the true model if enough data is observed. When the true model is not among those being considered, Bayesian model selection chooses the candidate that is closest to the true model in terms of Kullback-Leibler divergence [4, 5, 18]. In practice, the true model is typically not known and is potentially not in $\mathbf{M}$. Despite this, careful selection of a class of approximating models is important in the understanding of many problems. Here we seek a model within the class that is approximately correct (containing only significant predictors) and that approximates the parameters of the true underlying response model with low variance [9].

Atkinson [1] raises several concerns about inference for regression models. Our prior probability framework avoids by design his concern about improper prior distributions for models. A concern about nesting, so that two models may be true, is resolved by noting that the simpler model will be identified as more data is collected, and that simpler models are more desirable explanations [19, 28]. Atkinson [1] also indicates that if two response models are compared, the true model and an incorrect model with fewer parameters, then asymptotically the correct model will be selected, but that for finite numbers of samples the posterior probabilities may support the incorrect model in the absence of strong evidence from the data. This is a cause for care, but is not a violation of the likelihood principal, and negative consequences for selecting a model when the data do not provide enough evidence is a problem for any selection criterion. A goodness-of-fit test may be useful to provide further *post hoc* validation.

2.2.3 Entropy-Based Criteria

Several authors [6, 16, 26] proposed the use of the expected gain in Shannon information (or decrease in entropy) given by an experiment as an optimal design criterion. This expected gain is a natural measure of the utility of an experiment. The choice of design influences the expected gain in information as the predictive distribution of future output $\mathbf{Z}$ is determined by the design $\mathbf{x}$, model M_ℓ, and the prior distribution in Eq. (2)

$$p(\mathbf{Z} \mid M_\ell, \sigma^2, \mathbf{x}) \sim \text{Normal}\left(\mathbf{y}\boldsymbol{\mu}_\ell, \sigma^2\left[\mathbf{y}V_\ell\mathbf{y}^{\mathrm{T}} + I_n\right]\right). \tag{3}$$

The marginal distribution of $\mathbf{Z}$ given $M_\ell, \mathbf{y}$, obtained by integrating out σ^2, is a multivariate t distribution. Entropy is different for discrete (model selection) and continuous (parameter estimation) random variables, so each is discussed in turn.

For model selection, Box and Hill [10] use the expected increase in Shannon information J as a design criterion. The criterion was derived from information theory where the information (entropy) was used as a measure of uncertainty for distinguishing the s candidate models.

$$\begin{aligned} J &= -\sum_{\ell=1}^{s} p(M_\ell) \log p(M_\ell) + \int \left(\sum_{\ell=1}^{s} p(M_\ell \mid \mathbf{Z}, \mathbf{y}_\ell) \log p(M_\ell \mid \mathbf{Z}, \mathbf{y}_\ell) \right) p(\mathbf{Z} \mid \mathbf{y}_\ell) d\mathbf{Z} \quad (4) \\ &= \sum_{\ell=1}^{s} p(M_\ell) \int \log \frac{p(\mathbf{Z} \mid M_\ell, \mathbf{y}_\ell)}{\sum_{l=1}^{s} p(\mathbf{Z} \mid M_l, \mathbf{y}_\ell) p(M_l)} p(\mathbf{Z} \mid M_\ell, \mathbf{y}_\ell) d\mathbf{Z} \end{aligned}$$

An explicit solution is unknown in general, so J may be evaluated numerically or approximated. Alternately, Box and Hill [10] gave an upper bound approximation, the expected gain in Shannon information between the predictive distributions of each pair of candidate models M_i and M_l. This approximation was originally named the D-criterion, but we use the notation MD, as in [28].

$$MD = \sum_{0 \le i \ne l \le s} p(M_i) p(M_l) \left(\int p(\mathbf{Z} \mid M_i, \mathbf{y}_i) \log \frac{p(\mathbf{Z} \mid M_i, \mathbf{y}_i)}{p(\mathbf{Z} \mid M_l, \mathbf{y}_l)} d\mathbf{Z} \right) \quad (5)$$

The MD criterion is effective in practice and popular with research workers [21]. We use MD for the model discrimination portion of our joint criterion. For the normal linear model, Meyer et al. [28] show that MD reduces to a closed form if a noninformative prior $1/\sigma$ on σ and a conditionally normal prior for $\boldsymbol{\beta}$ given σ are assumed. A closed form also results if the conjugate prior is assumed.

Proposition 1. *Assume the conjugate normal gamma prior in Eq. (2). Let* $\hat{\mathbf{z}}_\ell = \mathbf{y}_\ell \boldsymbol{\mu}_\ell$, *and* $\boldsymbol{V}_\ell^* = \left[\mathbf{y}_\ell \boldsymbol{V}_\ell \mathbf{y}_\ell^{\mathrm{T}} + I \right]$. *Then for the linear model,* MD *simplifies to*

$$MD = \sum_{0 \le i \ne l \le s} \frac{1}{2} p(M_i) p(M_l) \left[-n + tr(\boldsymbol{V}_l^{*-1} \boldsymbol{V}_i^*) + \frac{1}{\lambda} \left(\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l \right)^{\mathrm{T}} \boldsymbol{V}_l^{*-1} \left(\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l \right) \right] \quad (6)$$

Proof. See Appendix A.1 □

For parameter estimation, Bernardo [6] and Smith and Verdinelli [38] adopted an entropy based method to ensure precise estimates for parameters that have already been identified as important. They choose the design that maximizes the expected gain in Shannon information (or equivalently,

maximizes the expected Kullback-Leibler distance) between the posterior and prior distributions of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

$$BD = \int\int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \left[\frac{p(\boldsymbol{\theta} \mid \mathbf{Z})}{p(\boldsymbol{\theta})} \right] d\boldsymbol{\theta} d\mathbf{Z} \tag{7}$$

Eq. (7) simplifies considerably for the normal linear model into a form known as the Bayesian D-optimal criterion (hence the choice of name BD).

Proposition 2. *For a linear model M_ℓ of the form Eq. (1), the prior probability model Eq. (2), and a given design $\mathbf{y}_\ell$,*

$$BD = \frac{1}{2} \log \left| \mathbf{y}_\ell^{\mathrm{T}} \mathbf{y}_\ell + \boldsymbol{V}_\ell^{-1} \right| - \frac{1}{2} \log \left| \boldsymbol{V}_\ell^{-1} \right|$$

Proof. See Appendix A.2. □

Following Borth [8], the entropy criterion S_P for parameter uncertainty generalizes when there are multiple candidate models.

$$\begin{aligned} S_P &= -\sum_{\ell=1}^{s} p(M_\ell) \int p(\boldsymbol{\theta}_\ell \mid M_\ell) \log p(\boldsymbol{\theta}_\ell \mid M_\ell) d\boldsymbol{\theta}_\ell \\ &\quad + \int \sum_{\ell=1}^{s} p(M_\ell \mid \mathbf{Z}) \int p(\boldsymbol{\theta}_\ell \mid \mathbf{Z}, M_\ell) \log p(\boldsymbol{\theta}_\ell \mid \mathbf{Z}, M_\ell) d\boldsymbol{\theta}_\ell \sum_{l=1}^{s} p(M_l) p(\mathbf{Z} \mid M_l) d\mathbf{Z} \end{aligned} \tag{8}$$

Proposition 3. *For the normal linear model, S_P simplifies to*

$$S_P = \sum_{\ell=1}^{s} \frac{p(M_\ell)}{2} \log \left| \mathbf{y}_\ell^{\mathrm{T}} \mathbf{y}_\ell + \boldsymbol{V}_\ell^{-1} \right| + K \tag{9}$$

for some K that does not depend on the design.

Proof. See Appendix A.3. □

2.2.4 Joint Criterion

In order to account for both model discrimination and parameter estimation simultaneously, Hill et al. [22] proposed a joint criterion that adds a weighted measure of discrimination and precision,

$$C = w_1 D_0 + w_2 E_0, \tag{10}$$

where D_0 is some measure of discrimination and E_0 is some measure of precision in parameter estimation. A nonunique choice of D_0 and E_0 they suggest is the model discrimination criterion

proposed by Box and Hill [10], and the determinant of the regression matrix for estimating the parameters for model i.

$$C = w\frac{MD}{MD^*} + (1-w)\sum_{i=1}^{s} p(M_i)\frac{E_i}{E_i^*}, \tag{11}$$

where MD^* and E_i^* are the maximum values of MD and E_i over the design region, and w is a nonnegative weight placed on model discrimination. They assumed that σ^2 is known or can be estimated when computing C. As the two criteria are summed together and weighted by w, the maxima MD^* and E_i^* may be less relevant than the range of the criteria over the design space.

Borth [8] treated the two objectives using the idea of the change in total entropy. He showed that it decomposed into the model discrimination term J and parameter estimation term S_P. We denote Borth's criterion as B hereafter. The scale for entropy for continuous random variables (parameters) may not be well-calibrated with entropy for a discrete random variable (model selection): their range may differ when evaluated throughout the design space. Borth's method also requires computationally expensive numerical integration.

Here we also use the idea of the expected gain in entropy of an experiment, but normalize J and S_P over their range of values, and simplify the weight factor. Instead of numerically evaluating the J criterion, we approximate it with the MD criterion. So an upper bound approximation of the joint criterion for model discrimination and parameter estimation is

$$S_Q = w\frac{MD - MD_{min}}{MD_{max} - MD_{min}} + (1-w)\frac{S_P - S_{P_{min}}}{S_{P_{max}} - S_{P_{min}}}, \tag{12}$$

where $MD_{min}, MD_{max}, S_{P_{min}}, S_{P_{max}}$ are the smallest and largest MD values and the smallest and largest S_P values respectively over all designs in $\mathbf{D}$, and $w \in [0,1]$ is a weight factor. This is similar in form to criterion C, but reduces to a closed form if the prior setup in Eq. (2) is used, as a result of Eq. (6) and Eq. (9). Eq. (12) does not require σ^2 to be known (see the propositions above for linear models, and comments below for nonlinear models), and incorporates prior information and data from initial experiments.

The weight w should be selected based on the results of the initial experiments and the focus of the follow-up experiment. If the initial experiment was insufficient to identify the important parameters, then more weight should be placed on model discrimination. If the model is reasonably determined, then more focus can be placed on parameter estimation. Hill et al. [22] suggested $w = \left[\frac{s}{s-1}\left(1 - p(M_{max})\right)\right]^{\xi}$ where M_{max} is the *a priori* most probable model. Another choice is $w = [(1 - (p(M_{max}) - p(M_{max2}))]^{\xi}$, where M_{max2} is the second most probable model. Small values of ξ places more weight on model discrimination. To equally balance the two calibrated

entropy measures, w can also be set at $1/2$. The numerical examples in Section 3 use both the weighting function of Hill et al. [22] and $w = 1/2$. Examples 1 and 2 assess the dependence upon the optimal design on the weight. The examples show that rescaling the entropies can be important, but that the final design may be somewhat insensitive to a misspecification in w.

To achieve the joint objectives of model discrimination and parameter estimation, we seek a design $\mathbf{x} \in \mathbf{D}$ that maximizes S_Q in Eq. (12). For normal linear models, S_Q simplifies to a closed form through Eq. (6) and Eq. (9). The criterion is also applicable to nonlinear models. When a nonlinear model can be approximated by a linear model in the neighborhood of θ_0, S_Q can be applied by substituting the initial estimates of the parameters [12]. For non-normal models, S_Q requires numerically integrating Eq. (5) and Eq. (7). For generalized linear and nonlinear models, Bayesian methods [3, 17] can be used to approximate the terms in Eq. (5) and Eq. (7).

Shannon information is not the only possible approach to develop a joint model selection and parameter estimation design criterion. Bingham and Chipman [7] propose a weighted average of Hellinger distances between predictive densities of all possible pairs of competing models as a criterion for model discrimination. A linear combination of Bingham and Chipman [7]'s criterion and a weighted average of the Hellinger distances between the prior and posterior distributions of each model's parameters can also be used as a joint criterion. For the prior setup in Eq. (2), this reduces to a closed form. The weighting functions described above can be used to weight the importance of each objective. The upper bound on the Hellinger distances for each individual term can be useful for rescaling, but the maximum values of each term for a particular finite n, can be quite far from the upper bound and rescaling each term by its upper bound may not be appropriate. We do not consider that combined criterion further here.

2.3 Some Computational Issues

Although S_Q simplifies to a closed form for the normal linear model, there are computational challenges. We consider three here. First, the number of models grows exponentially in the number of predictors. Second, the min and max values of the two entropy measures that comprise S_Q are required. Third, the number of designs grows combinatorially in the number of candidate runs.

To address the first issue, the summands for MD and S_P are computed by using only the most likely models. There are typically far fewer than $s = 2^p$ different models whose probability $p(M_\ell)$ lead it to be a competitor for the 'best' after the initial stage of experimentation. By considering only the most likely models, Eq. (12) becomes tractable. There are several ways one can chose a subset of probable models: (i) Pick all models h so that $p(M_h) \geq E$, (ii) Pick the h most likely

models, where h is the smallest integer so that $p(M_{(1)}) + p(M_{(2)}) + \cdots + p(M_{(h)}) \geq F$. Raftery et al. [34] take a similar approach to model averaging. Examples 2 and 3 of Section 3 use (i) with $E = 0.02$. The top models have higher posterior probabilities, so we set $E = 0.02$. In Section 3.1, no model clearly stands out after the initial runs, so we use $E = 0.008$.

When direct enumeration is not computationally feasible, these 'more important' models can be identified heuristically by using Markov Chain Monte Carlo methods like MC^3 (Markov Chain Monte Carlo Model Composition) [27] to estimate the $p(M_\ell)$. The state space for MC^3 is the set of s models, and a sample path visits a sequence of different models, M_ℓ. Candidate states for transitions are chosen from the set of models with one more or one fewer active predictors. The relative probabilities for the current and candidate states, needed to implement the Metropolis-Hastings step of MC^3, can be computed from closed-form formulas in Raftery et al. [34]. The number of times a model is visited during MC^3 divided by the number of iterations of MC^3 is a consistent estimate of the model's posterior probability. Chipman et al. [15] and Ng [31] discuss some practicalities of Markov chain Monte Carlo methods for model selection.

Second, we use an optimization heuristic to estimate $MD_{min}, MD_{max}, S_{P_{min}}, S_{P_{max}}$. We use the k-exchange algorithm of Johnson and Nachtsheim [23] to search for the maximum and minimum values. The k-exchange algorithm was first proposed to construct D-optimal designs, but because it is a general algorithm, it can be used to select from a finite set of designs as long as an optimality criterion is given. Numerical results [23, 33] show that it is efficient and effective in constructing optimal designs, and the algorithm has been widely used. In the numerical examples we considered, the k-exchange algorithm was very efficient in identifying the optimal designs. In addition to increasing k as suggested in [23], we also found that for the k-exchange algorithm, increasing the number of starting designs from scattered points in the design space improves the search for the optimal. Alternatively, the branch and bound algorithm in [39], or nested partitions [37] can be used to find the global optima.

Third, we generalize the k-exchange algorithm (Appendix A.4) to identify a design with a high value of S_Q to improve the scaling of the entropy measures. The algorithm is a greedy algorithm that swaps in and out design points one at a time.

More work on computational issues is an avenue for future research.

3. Numerical Results

Three numerical experiments compare the new criterion, S_Q, with the two other joint criteria in the literature, as well as the MD and S_P criteria. The optimal S_Q follow-up design $\mathbf{x}^*(w, \mathbf{z}_0)$ depends upon the weight w and previous observations $\mathbf{z}_0$. Let $\mathbf{x}^*(\mathbf{z}_0) = \{\mathbf{x}^*(w, \mathbf{z}_0) : w \in [0, 1]\}$ be the set of designs that the S_Q criterion identifies, given $\mathbf{z}_0$.

3.1 Chemical Reactor Experiment

Box et al. [11, p. 377] gave data for a chemical-reactor experiment that used a 2^5 full factorial design. From this data, we extracted runs that correspond to five columns of a Plackett-Burman 12 run (PB12) design. We treated those runs (see Table 1) as an initial experiment. The follow-up design was simulated by extracting the remaining runs from the complete experiment.

Table 1: PB12 design and data extracted from the full 2^5 reactor experiment

Run i	A	B	C	D	E	z_i
1	1	1	-1	1	1	77
2	1	-1	1	1	1	42
3	-1	1	1	1	-1	95
4	1	1	1	-1	-1	61
5	1	1	-1	-1	-1	61
6	1	-1	-1	-1	1	63
7	-1	-1	-1	1	-1	69
8	-1	-1	1	-1	1	59
9	-1	1	-1	1	1	78
10	1	-1	1	1	-1	60
11	-1	1	1	-1	1	67
12	-1	-1	-1	-1	-1	61

We considered fifteen predictors (five factors and their two factor interactions) to get 2^{15} distinct linear models in the model space $\mathbf{M}$, each differing by the absence or presence of each predictor. We used the equal probability prior for model uncertainty, $p(M_\ell) = 2^{-15}$, and the prior for parameters suggested by Raftery et al. [34]. Table 2 shows the probabilities for the top 8 models, given that prior distribution and the PB12 data. No model clearly stands out, but the model identified in the original analysis of all 32 runs [11], with factors (B, D, E, BD, DE), is ranked best.

To distinguish between the top eight models, $n = 3$ additional runs were selected from the remaining 20 runs. The best designs for each joint criterion (S_Q with $w = 0.5$; B; and C with $\xi = 2$ as in [22]) were computed by evaluating the criteria over each possible design. The joint

Table 2: Probability of the eight most probable models after 12 runs

Model	Posterior Probability
B, D, E, BD, DE	0.0483
B, C, D, E, BD, DE	0.0206
B, D, E, BC, BD, DE	0.0195
B, D, E, BD, BE, DE	0.0106
A, B, D, E, BD, DE	0.0105
B, D, E, AE, BD, DE	0.0101
B, D, E, AC, BD, DE	0.0085
B, D, E, BD, CD, DE	0.0083

Table 3: Posterior probability (Post.) of the three most probable models with PB12, + 3 runs determined by the best design obtained from full enumeration of the S_Q, B, and C criteria.

New S_Q Criterion		Borth's B Criterion		Hill's C Criterion	
Model	Post.	Model	Post.	Model	Post.
B, D, E, BD, DE	0.189	B, D, E, BD, DE	0.132	B, D, E, BD, DE	0.166
B, D, E, BC, BD, DE	0.053	B, C, D, E, BD, DE	0.082	B, D, E, BC, BD, DE	0.065
A, B, D, E, BD, DE	0.045	A, B, D, E, BD, DE	0.051	B, D, E, BD, BE, DE	0.04

criterion S_Q results in different designs than the B and C criteria. The posterior probabilities of all models were then recomputed using all 15 runs, and the top 3 models are shown in the left portion of Table 3. All three designs identified the same top model identified in the original analysis of all 32 runs. S_Q discriminated in favor of the top model more than criteria B and C. Table 4 indicates that S_Q reduced the parameter generalized variance (the determinant of the posterior covariance matrix of the parameter estimates, $|V(\boldsymbol{\beta})|$) of the top model more than B and C.

To compare the computational burden, each criterion was evaluated for all possible designs for one, two, three and four additional runs using Maple8 (slow, because it is interpreted, but relative CPU times are illustrative). Table 5 shows the computation times for the S_Q and B criterion. The computation times for S_Q and C were similar. The curse of dimensionality made quadrature an inefficient approach for the numerical integrations required by B.

Table 6 shows the posterior probabilities of the top three models with the model discrimina-

Table 4: Parameter generalized variance $|V(\boldsymbol{\beta})|$ for the *a posteriori* top model (B, D, E, BD, DE), given PB12 + 3 runs, based on the S_Q, B and C criteria.

Criterion	$\lvert V(\boldsymbol{\beta})\rvert$
S_Q	3.62×10^{-12}
B	4.81×10^{-12}
C	4.74×10^{-12}

Table 5: CPU time for computing S_Q, C and B (hours).

Additional runs	S_Q and C	B
1	0.005	0.007
2	0.04	0.86
3	0.33	55.4
4	1.87	453

Table 6: Posterior probability (Post.) of the three most probable models with PB12 + 3 runs determined by the best design obtained by full enumeration for S_Q, MD, and S_P.

S_Q Criterion		MD Criterion		S_P Criterion	
Model	Post.	Model	Post.	Model	Post.
B, D, E, BD, DE	0.189	B, D, E, BD, DE	0.156	B, D, E, BD, DE	0.124
B, D, E, BC, BD, DE	0.053	B, D, E, BD, BE, DE	0.048	B, C, D, E, BD, DE	0.072
A, B, D, E, BD, DE	0.045	B, C, D, E, BD, DE	0.038	A, B, D, E, BD, DE	0.032

tion MD and parameter estimation S_P criteria. As expected, MD did a better job than S_P at distinguishing the top model from the others, and Table 7 indicates that S_P outperformed MD at reducing the parameter generalized variance of the top model. S_Q on the other hand outperformed MD in favoring the top model, and was only slightly poorer than S_P in parameter estimation.

This example also illustrates the importance of normalizing that we suggest, as one of the subcriteria would be ignored without recalibration. With the equal probability prior for each model, the range of uncalibrated MD scores over the design space ranged from 0.034 to 0.062, while the uncalibrated S_P scores range from 0.88 to 0.90. Without recalibration, the joint criterion would have selected the best S_P design, and ignored the model discrimination objective.

We tested the sensitivity to the prior distribution by rerunning the experiment with the independence prior $p(M_\ell) = \omega^{t_\ell}(1-\omega)^{p-t_\ell}$, where t_ℓ is the number of predictors in model ℓ, ($\ell = 1, \ldots, 2^{15}$), with $\omega = 0.25$. The model with factors (B, D, E, BD, DE) is ranked third when only 12 runs are used, but the S_Q criterion again identified (B, D, E, BD, DE) as the most probable model after the $n = 3$ run follow up was completed.

To test the sensitivity of the designs to the weights, w was varied from 0 to 1. In this example,

Table 7: Parameter generalized variance $|V(\boldsymbol{\beta})|$ for the *a posteriori* top model (B, D, E, BD, DE), given PB12 + 3 runs, based on the S_Q, MD and S_P criteria.

Criterion	$\lvert V(\boldsymbol{\beta})\rvert$
S_Q	3.62×10^{-12}
MD	4.74×10^{-12}
S_P	2.13×10^{-12}

the top S_Q design is robust over a range of weights. There were only three different top designs obtained as w was varied from 0 to 1. When $0 \leq w < .27$, the top S_P design is obtained. The same design obtained for S_Q when $w = 0.5$ is obtained when $0.27 \leq w < 0.78$. The top MD design is obtained when $w \geq .78$. For the weighting function suggested in [22], the best MD design is selected when ξ gets small, $\xi \leq 4$, and the best S_P design is selected when ξ gets large, $\xi > 27$, and the same design selected for any ξ in between. In this example, small changes in w or ξ do not significantly change the optimal design.

The Entropy Balancing k-Exchange Algorithm in Appendix A.4 was implemented to evaluate its effectiveness to find good designs in this problem. The best design is known for this problem because an exhaustive evaluation of all 1140 designs is possible. The Entropy Balancing algorithm consists of two steps. In the initial step of the algorithm, the maximum and minimum values of MD and S_p are estimated using the k-Exchange$_{max}$ and k-Exchange$_{min}$ algorithms. In the second step, the k-Exchange$_{max}$ algorithm is used to search for a good S_Q design. We first evaluated how well the initial step performs in determining the maximum and minimum values. We varied the number r of initial random designs to start the algorithm, $r = 100$ and $r = 200$, and conducted 10 independent replications of the algorithm for each r (each replication samples an independent set of r design points). When $r = 100$, the estimates of ($MD_{max}, MD_{min}, S_{P_{min}}, S_{P_{max}}$) were all equal to the actual value 70% of the time. When $r = 200$, all four estimates were correct 80% of the time. In the remaining cases, only one of the four actual values was not obtained, but the estimate was close to the actual value. We next tested the how well the second step of the algorithm identifies the known best optimal S_Q design. The known best optimal S_Q design was identified 70% of the time with $r = 100$, and 90% of the time with $r = 200$. The remaining nonoptimal designs selected were among the top 8 designs.

3.2 Finding a Known Model

To determine how well the criteria performs in detecting a known model, and how the best S_Q model depends upon w in repeated samples, we ran 50 replications of an experiment using the S_Q criterion on a known model with 4 potential factors (A, B, C, D), namely $Z = 10A + 15B + 6AB + 7AC + \zeta_i$, where $\zeta_i \sim$ Normal$(0, 5)$. For each replication i, we ran an initial 2^{4-1} fractional factorial design to generate preliminary data $\mathbf{z}_{0,i}$, which was then used to create a prior distribution for a follow-up design with $n = 3$ runs as described in Fig. 1. In each replication, the true (known) model was among the top eight candidate models after the initial $2^{4-1} = 8$ runs but was completely confounded with three other models. The three additional runs selected by S_Q dealiased

Figure 1: Algorithm to assess the identification of a known model in Section 3.2

For $i = 1, 2, \ldots, 50$:
1. Generate independent preliminary data $\mathbf{z}_{0,i}$ with a 2^{4-1} factorial design.
2. Update the distributions for the unknown models and parameters as in Section 2.2.1.
3. Determine the best S_Q design $\mathbf{x}^*(w, \mathbf{z}_{0,i})$ for $n = 3$ additional runs, as a function of w.
4. Run the best follow-up design, $\mathbf{x}^*(w, \mathbf{z}_{0,i})$, for each w.
5. Compute the posterior probability and ordinal rank of each model, as a function of w.

End for loop

the confounded effects and distinguished between the top competing models.

In each replication, three different designs were obtained ($|\mathbf{x}^*(\mathbf{z}_{0,i})| = 3$) as w varied between 0 and 1. The best S_Q design $\mathbf{x}^*(w, \mathbf{z}_{0,i})$ was the top S_P design ($\mathbf{x}^*(0, \mathbf{z}_{0,i})$) for small w. For a certain range between 0 and 1, a unique top S_Q design was obtained that balances model discrimination and parameter estimation. For larger w, the best S_Q design was the best MD design ($\mathbf{x}^*(1, \mathbf{z}_{0,i})$). In 70% of the replications, the same three designs $\mathbf{x}^*(\mathbf{z}_{0,i})$ were selected and the same S_Q was selected for w in approximately the range (0.1, 0.55). The other 30% of the replications resulted in 3 other sets of S_Q designs with the same qualitative features: the designs with w in the range of about 0.1 up to 0.4-0.8 balanced model discrimination and parameter estimation.

In 49 out of 50 replications, the true model was identified as the best model when the S_Q design with intermediate values of w was used to balance discrimination and estimation. In the remaining replication, the true model had the third highest posterior probability. Averaging over 50 replications, the probability that the true model was best improved from 0.04 after the initial stage (8 runs), to 0.21 (after the 3 follow-up runs). The average range of MD was 7.42, the average range of S_P was 0.59, and $MD_{max} > S_{P_{max}}$ for all replications. If the individual measures were not recalibrated by their ranges, the S_Q criterion would have selected the best MD design and ignored parameter estimation unless a very small weight were placed on model discrimination.

This experiment gave the same qualitative conclusions as Section 3.1. Borth's criterion took orders of magnitude more time to compute due to numerical integration issues (curse of dimensionality). Rebalancing the entropy measures was important for assuring a balance between discrimination and estimation. The optimal design was not highly sensitive to the choice of w.

3.3 Critical Care Facility

The critical care facility illustrated in Fig. 2 was originally studied by Schruben and Margolin [36]. Patients arrive according to a Poisson process and are routed through the system depending

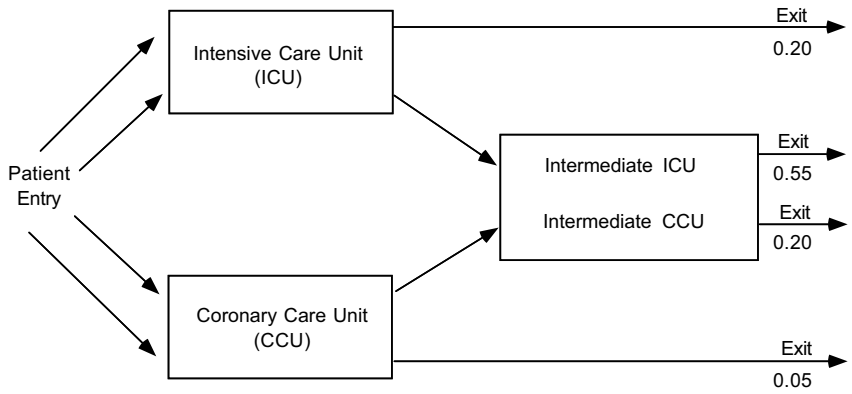

Figure 2: Estimated fraction of patients routed through the units of a critical care facility.

upon their specific health condition. Stays in the intensive care (ICU), coronary care (CCU), and intermediate care facilities are presumed to be lognormally distributed. This section compares S_Q with C and the individual criteria, MD and S_P. Borth's criterion B took too much time to evaluate and was therefore not compared. We initially ran a 64 run design using the Bayesian model average to sample uncertain input parameters [32]. We considered twelve input parameters, resulting in 2^{12} distinct linear models in the model space $\mathbf{M}$, each differing by the absence and presence of each predictor. Table 8 shows the posterior probabilities for the top 5 models. The model identified in a 128 run study in [32] is ranked fifth here.

Schruben and Margolin [36] studied how to allocate random number streams to reduce variability in response surface parameter estimates. Their response model predicts the expected number of patients per month $E[Z]$ that are denied entry to the facility as a function of the number of beds in the ICU, CCU, and intermediate care facilities. They presume fixed point estimates for $k = 6$ input parameters, one per source of randomness, to describe the patient arrival process (Poisson arrivals, mean $\hat{\lambda} = 3.3$/day), ICU stay duration (lognormal, mean 3.4 and standard deviation 3.5 days), intermediate ICU stay duration (lognormal, mean 15.0, standard deviation 7.0), intermediate CCU stay duration (lognormal, mean 17.0, standard deviation 3.0), CCU stay duration (lognormal, mean 3.8, standard deviation 1.6), and routing probabilities (multinomial, $\hat{p_1} = 0.2$, $\hat{p_3} = 0.2$, $\hat{p_4} = 0.05$). Some parameters are multivariate, and there are a total of $1 + 4 * 2 + 3 = 12$ dimensions of parameters. For the lognormal service times, the log of the service times has mean μ and precision $\lambda = 1/\sigma^2$. Subscripts distinguish the parameters of each service type (e.g., $\mu_{\text{icu}}, \mu_{\text{iicu}}$, $\mu_{\text{iccu}}, \mu_{\text{ccu}}$, λ_{icu}). The analysis here presumes a linear response model in these 12 parameters.

The actual system parameters are not known with certainty, and the estimated system performance will be in error if the actual parameter values differ from their point estimates. As in Ng and

Table 8: The five most probable models after 64 runs.

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.43298
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_3, p_4$	0.20019
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \lambda_{\text{iccu}}, p_1, p_4$	0.0682
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_4$	0.0516
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.0296

Table 9: The most probable models with 64+32 runs determined by S_Q with $w = 0.55$.

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.407
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_3, p_4$	0.215
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.110

Chick [32], who used naive Monte Carlo sampling for unknown inputs to do an uncertainty analysis, we fix the number of beds in each of the three units (14 in ICU, 5 in CCU, 16 in intermediate care), and study how the expected number of patients per month that are denied entry depends on the unknown parameters. Design points for the unknown parameter values could take on values of the MLE $\pm$ one standard error. The approach of Raftery et al. [34] was used to obtain prior distributions for the unknown response parameters.

We used the S_Q criterion with $w = 0.55$ (or $\xi = 1$) to avoid focusing on parameter estimation too early. The design points of a full factorial for the 12 parameters were candidates for the 32 run follow-up design. The number of possible 32 run designs from the 2^{12} candidate runs is large, we used the k-exchange algorithm to search for the best S_Q design ($r = 50$, $k = 5$), then ran the critical care simulations again with that design. The posterior probabilities for the top three models, given the data from the combined design (64+32), are shown in Table 9. The top model is the same model identified in the 128 runs analysis in [32], but the S_Q criterion identified this model with fewer runs. We also used the k-exchange algorithm ($r = 50$ and $k = 5$) to determine a good C design. Table 10 shows the posterior probabilities after running the simulations with the C design. The C design identified the same model as S_Q, but S_Q did slightly better in discriminating the top two models.

The best designs for MD and S_P are different than the best S_Q design with $w = 0.55$. The MD design identified the same top model as the S_Q design, and discriminated between the top two models slightly better than the S_Q design (Table 9 and Table 11). Table 12 indicates that design S_Q did a better job than C and MD at reducing the parameter generalized variance of the top model. The S_Q criterion with $w = 0.75$ (or $\xi = 0.5$) resulted in better model discrimination than with

Table 10: Most probable models with 64+32 runs with the C criterion.

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.359
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_3, p_4$	0.232
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \lambda_{\text{iccu}}, p_1, p_3, p_4$	0.077

Table 11: Most probable models with 64+32 runs with the MD criterion.

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.438
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, \lambda_{\text{iccu}}, p_1, p_3, p_4$	0.125
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.109

$w = 0.55$ at the cost of slightly less effective parameter estimation (in this case, $|\mathbf{x}^*(\mathbf{z}_{0,i})| > 3$).

The top two models identified in the S_P design were the same models identified in the original 64 run analysis, and the top model identified by the S_Q and MD design is only ranked fourth when the S_P design is used. The S_P criterion focused on designs that had good parameter estimation primarily for models with higher posterior probability. Using the S_P criterion too early in the experimentation process can prematurely focus the design and experimentation on a few models that may or may not be good approximations to the system (because of the small number of runs), an issue raised by Atkinson [1]. An early focus on MD can better distinguish competitors for the best model, but at the expense of poorer parameter estimates. S_Q balanced both of those needs.

4. Discussion and Conclusions

The purpose of many experiments is to distinguish between likely mathematical models and obtain precise estimates for the model parameters. The three joint design criteria examined here each use an additive measure for entropy measures or bounds for model and parameter uncertainty. Our

Table 12: Parameter generalized variance $|V(\boldsymbol{\beta})|$ after 64+32 runs for the model with $\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$.

Runs	Criterion	$\lvert V(\boldsymbol{\beta})\rvert$
96	S_P	4.63×10^{-28}
96	$S_{Q_{w=0.55}}$	4.69×10^{-28}
96	$S_{Q_{w=0.75}}$	5.33×10^{-28}
96	C	5.70×10^{-28}
96	MD	5.71×10^{-28}
64	–	9.46×10^{-26}

proposal for the new S_Q criterion to normalize each entropy measure by the amount each varies over the design space provides an insight that the other joint criteria do not: It indicates how rich the design space is for improving each entropy measure. If the range for one of the component criteria is much smaller than for the other (e.g. $MD_{max} - MD_{min} >> S_{P_{min}} - S_{P_{max}}$), or if the number of potentially optimal designs, $|\mathbf{x}^*(\mathbf{z}_0)|$, is small, then a richer design space might be considered.

S_Q is computationally more efficient than Borth's joint criterion especially when the planned follow-up designs get larger. In the first two examples, the S_Q design performs as well as Borth's criterion, but it is computationally more efficient and practical. S_Q extends the C criterion as it considers the relevant range of the individual criteria, does not require initial estimates of the variance, and accounts for available prior information. These two examples also show that there are 3 different designs for S_Q as w varies from 0 to 1. The optimum MD design is selected when the model discrimination term is heavily weighted and the optimum S_p design is selected when the parameter estimation term is heavily weighted. For each experiment, the S_Q design that balances both objectives is shown to be insensitive to a range of w, and this best design selected performs more efficiently than the other criteria.

Three numerical experiments illustrated the compromise between model discrimination and parameter estimation obtained when using the joint criterion S_Q. Compared with the individual criteria, the balanced S_Q design was about as good as the MD design for model discrimination, and was almost as good as the S_P design for parameter estimation. The MD design fared less well for parameter estimation, and the S_P design was least effective for model discrimination.

Although S_Q is easier to compute for the linear model than Borth's criterion, the large number of matrix calculations required to compute the S_Q criterion may need to be balanced against the cost of actually running the experiments. In a simulation context, CPU cycles might be better spent running replications rather than computing S_Q if the simulations run quickly. For expensive industrial experiments or complex simulations with long run times, the S_Q criterion may be an effective mechanism to balance the needs of factor identification and parameter estimation.

Sequential designs and criteria based on the Hellinger distance are avenues for further research.

A. Mathematical Details

A.1 Proof of Prop. 1

Conditioning on σ^2, the MD criterion can be rewritten

$$MD = \sum_{0 \le i \ne l \le s} p(M_i)p(M_l) \int_0^\infty p(\sigma^2) \int_{-\infty}^\infty p(\mathbf{Z} \mid M_i, \mathbf{y}_i, \sigma^2) \log \frac{p(\mathbf{Z} \mid M_i, \mathbf{y}_i, \sigma^2)}{p(\mathbf{Z} \mid M_l, \mathbf{y}_l, \sigma^2)} d\mathbf{Z} d\sigma^2 \quad (13)$$

Meyer et al. [28] substituted the predictive distribution of the normal form in Eq. (3) into Eq. (13) and integrated with respect to $p(\mathbf{Z} \mid M_i, \sigma^2)$ to obtain:

$$\begin{aligned} MD &= \sum_{0 \le i \ne l \le s} p(M_i)p(M_l) \cdot \left[\int_0^\infty \pi(\sigma^2) \left[\frac{1}{2} \log \left(\frac{|\mathbf{V}_l^*|}{|\mathbf{V}_i^*|} \right) - \frac{1}{2\sigma^2} \right. \right. \\ &\quad \left. \left. \times \left(n\sigma^2 - \sigma^2 \text{tr}(\mathbf{V}_l^{*-1}\mathbf{V}_i^*) - (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l)^{\mathrm{T}} \mathbf{V}_l^{*-1} (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l) \right) \right] d\sigma^2 \right] \end{aligned}$$

We now isolate the dependence on the noninformative prior.

$$\begin{aligned} MD &= \sum_{0 \le i \ne l \le s} p(M_i)p(M_l) \cdot \left[\frac{1}{2} \log \left(\frac{|\mathbf{V}_l^*|}{|\mathbf{V}_i^*|} \right) - \int_0^\infty \frac{1}{2\sigma^2} \right. \\ &\quad \left. \times \left(n\sigma^2 - \sigma^2 \text{tr} \left(\mathbf{V}_l^{*-1}\mathbf{V}_i^* \right) - (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l)^{\mathrm{T}} \mathbf{V}_l^{*-1} (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l) \right) \pi(\sigma^2) d\sigma^2 \right] \\ &= \sum_{0 \le i \ne l \le s} \frac{1}{2} p(M_i)p(M_l) \cdot \left[\log \left(\frac{|\mathbf{V}_l^*|}{|\mathbf{V}_i^*|} \right) - n + \text{tr}(\mathbf{V}_l^{*-1}\mathbf{V}_i^*) \right. \\ &\quad \left. + (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l)^{\mathrm{T}} \mathbf{V}_l^{*-1} (\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_l) \int_0^\infty \frac{1}{\sigma^2} \pi(\sigma^2) d\sigma^2 \right] \end{aligned} \quad (14)$$

The double sum means that pairs i, l can be matched to make the log terms cancel out. And $\int_0^\infty \frac{1}{\sigma^2}\pi(\sigma^2)d\sigma^2 = [(\nu/2)/(\nu\lambda/2)] \int_0^\infty \text{InvertedGamma}\left(\sigma^2 \mid \left(\frac{\nu}{2}+1\right), \frac{\nu\lambda}{2}\right) d\sigma^2 = 1/\lambda$. Substitute this into Eq. (14) to justify the claim in Eq. (6).

A.2 Proof of Prop. 2

Condition on model M_ℓ and let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

$$\begin{aligned} BD &= \int \int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \left[\frac{p(\boldsymbol{\theta} \mid \mathbf{Z})}{p(\boldsymbol{\theta})} \right] d\boldsymbol{\theta} d\mathbf{Z} \\ &= \int \int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \left[p(\boldsymbol{\theta} \mid \mathbf{Z}) \right] d\boldsymbol{\theta} d\mathbf{Z} - \int \int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \left[p(\boldsymbol{\theta}) \right] d\boldsymbol{\theta} d\mathbf{Z} \end{aligned}$$

The second term on the right hand side of this last equation simplifies using Fubini's theorem and Bayes' theorem, assuming the integrals exist, and is independent of the design.

$$\begin{aligned}\int\int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log\left[p(\boldsymbol{\theta})\right] d\boldsymbol{\theta} d\mathbf{Z} &= \int\int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log\left[p(\boldsymbol{\theta})\right] d\mathbf{Z} d\boldsymbol{\theta} \\ &= \int \log\left[p(\boldsymbol{\theta})\right] p(\boldsymbol{\theta}) \int p(\mathbf{Z} \mid \boldsymbol{\theta}) d\mathbf{Z} d\boldsymbol{\theta} \\ &= \int \log\left[p(\boldsymbol{\theta})\right] p(\boldsymbol{\theta}) d\boldsymbol{\theta}\end{aligned}$$

So

$$BD = \int\int p(\mathbf{Z})p(\boldsymbol{\theta} \mid \mathbf{Z}) \log\left[p(\boldsymbol{\theta} \mid \mathbf{Z})\right] d\boldsymbol{\theta} d\mathbf{Z} - \int \log\left[p(\boldsymbol{\theta})\right] p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{15}$$

Condition now on both M_ℓ and σ^2 and focus on the inner integral in the left hand term of the last equation. It is well known that the conditional distribution of $\boldsymbol{\beta}$ given $\mathbf{z}, \sigma^2$ is a normal distribution with variance $\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})$ (e.g. see [5]). Call the posterior mean μ'. First integrate out $\boldsymbol{\beta}$, with σ^2 handled after.

$$\begin{aligned}&\int p(\boldsymbol{\beta} \mid \mathbf{Z}, M_i, \sigma^2) \log p(\boldsymbol{\beta} \mid \mathbf{Z}, M_i, \sigma^2) d\boldsymbol{\beta} \\ = &\int \left|\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right|^{\frac{1}{2}} (2\pi)^{-\frac{t_i+1}{2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \mu'\right)^{\mathrm{T}} \left(\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right)\left(\boldsymbol{\beta} - \mu'\right)\right] \\ &\times \left[\frac{\log\left|\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right|}{2} - \frac{(t_i+1)\log 2\pi}{2} - \frac{\left(\boldsymbol{\beta} - \mu'\right)^{\mathrm{T}} \sigma^{-2}\left(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1}\right)\left(\boldsymbol{\beta} - \mu'\right)}{2}\right] d\boldsymbol{\beta} \\ = &\frac{\log\left|\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right|}{2} - \frac{(t_i+1)\log 2\pi}{2} - \frac{\left|\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right| \left|\sigma^{-2}(\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1})\right|^{-1}}{2} \\ = &-(t_i+1)\log\sigma + \frac{1}{2}\log\left|\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1}\right| - \frac{1}{2}(t_i+1)\log 2\pi - \frac{1}{2}.\end{aligned}$$

Similarly,

$$\int p(\boldsymbol{\beta} \mid M_i, \sigma^2) \log p(\boldsymbol{\beta} \mid M_i, \sigma^2) d\boldsymbol{\beta} = -(t_i+1)\log\sigma + \frac{1}{2}\log\left|\mathbf{V}_i^{-1}\right| - \frac{1}{2}(t_i+1)\log 2\pi - \frac{1}{2}.$$

The difference in BD therefore has several terms that cancel out, justifying the claimed relationship.

$$\begin{aligned}BD &= \int\int \frac{1}{2}\log\left|\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1}\right| - \frac{1}{2}\log\left|\mathbf{V}_i^{-1}\right| d\mathbf{Z} d\sigma^2 \\ &= \frac{1}{2}\left(\log\left|\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{V}_i^{-1}\right| - \log\left|\mathbf{V}_i^{-1}\right|\right)\end{aligned}$$

A.3 Proof of Prop. 3

Substitute $p(M_i \mid \mathbf{Z}) = p(\mathbf{Z} \mid M_i)p(M_i)/\sum_{j=1}^{s} p(\mathbf{Z} \mid M_j)p(M_j)$ into Eq. (8), to obtain

$$\begin{aligned}
S_P &= -\sum_{i=1}^{s} p(M_i) \int p(\boldsymbol{\theta}_i \mid M_i) \log p(\boldsymbol{\theta}_i \mid M_i) d\boldsymbol{\theta}_i \\
&\quad + \int \frac{\sum_{i=1}^{s} p(M_i)p(\mathbf{Z} \mid M_i)}{\sum_{j=1}^{s} p(M_j)p(\mathbf{Z} \mid M_j)} \\
&\quad \times \int p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) \log p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) d\boldsymbol{\theta}_i \sum_{j=1}^{s} p(M_j)p(\mathbf{Z} \mid M_j) d\mathbf{Z} \\
&= -\sum_{i=1}^{s} p(M_i) \int p(\boldsymbol{\theta}_i \mid M_i) \log p(\boldsymbol{\theta}_i \mid M_i) d\boldsymbol{\theta}_i \\
&\quad + \sum_{i=1}^{s} p(M_i) \int p(\mathbf{Z} \mid M_i) \int p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) \log p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) d\boldsymbol{\theta}_i d\mathbf{Z} \\
&= \sum_{i=1}^{s} p(M_i) \Bigg[- \int p(\boldsymbol{\theta}_i \mid M_i) \log p(\boldsymbol{\theta}_i \mid M_i) d\boldsymbol{\theta}_i \\
&\quad + \int p(\mathbf{Z} \mid M_i) \int p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) \log p(\boldsymbol{\theta}_i \mid \mathbf{Z}, M_i) d\boldsymbol{\theta}_i d\mathbf{Z} \Bigg]
\end{aligned}$$

By Prop. 2,

$$\begin{aligned}
S_P &= \sum_{i=1}^{s} p(M_i) \left[\frac{1}{2} \log \left| \mathbf{y}_i^{\mathrm{T}} \mathbf{y}_i + \mathbf{V}_i^{-1} \right| - \frac{1}{2} \log \left| \mathbf{V}_i^{-1} \right| \right] \\
&= \sum_{i=1}^{s} \frac{p(M_i)}{2} \log \left| \mathbf{y}_i^{\mathrm{T}} \mathbf{y}_i + \mathbf{V}_i^{-1} \right| + K
\end{aligned}$$

where K is independent of $\mathbf{x}$, as claimed.

A.4 Variant of k-Exchange Algorithm

In the following, let x_α be the α^{th} row of the design matrix $\mathbf{x}$, and let a prime ($'$) denote an estimate of the primed variable. Define

$$S_{Q'} = \frac{MD - MD_{min'}}{MD_{max'} - MD_{min'}} + \frac{S_P - S_{P_{min'}}}{S_{P_{max'}} - S_{P_{min'}}}.$$

Let $S_{Q'}(x_\alpha)$ be $S_{Q'}$ evaluated with the design matrix $\mathbf{x}(\alpha)$, with row x_α removed from $\mathbf{x}$. A generic maximum k-exchange algorithm for an arbitrary criterion C_r follows. An algorithm for the minimum is similar.

k-Exchange$_{max}$ Algorithm

1. Order the n design points according to their $C_{r'}(x)$ values, $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, so that $\alpha < \beta$ implies $C_{r'}(x_\alpha) \leq C_{r'}(x_\beta)$.

2. For $\alpha = 1$ to k: delete $x_{(\alpha)}$ from design matrix **x**, add x^* from the list of all available design points where x^* maximizes $C_{r'}$ when added to design $\mathbf{x}(\alpha)$.

3. Repeat (reorder and replace points) until no improvement in $C_{r'}$ can be found.

Entropy Balancing k Exchange Algorithm

1. Estimate the range of entropy scores in order to calibrate them.

   (a) Select r random n run designs.

   (b) Estimate $MD_{max'}$

      i. For each of the r designs, implement a k-exchange$_{max}$ for criterion MD.

      ii. Let $\mathbf{R} \in \mathbf{D}$ be the set of designs obtained from 1(b)i. Set $MD_{max'} = \max_{d \in \mathbf{R}}(MD)$

   (c) Similarly estimate $S_{P_{max'}}$, $MD_{min'}$, $S_{P_{min'}}$.

2. For each of the r designs, implement the k-exchange$_{max}$ algorithm for criterion S_Q'

The algorithm does not guarantee optimality, but applying Steps 1 and 2 to each of the r randomly selected designs provides some protection against getting stuck in a local extrema.

For faster execution of the algorithm, k should be chosen small. However, Johnson and Nachtsheim [23] noted that a larger value of k leads to greater D-efficiency. A good choice of k to search for good S_Q designs depends on the number of predictors, the number of starting random designs r, and the number of additional runs n.

References

[1] Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika 65*(1), 39–48.

[2] Barton, R. R. (1998). Simulation metamodels. In D. J. Madeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan (Eds.), *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 167–174. Institute of Electrical and Electronics Engineers, Inc.

[3] Bennett, J.E., A. Racine. and J.C. Wakefield (1996). MCMC for nonlinear hierarchical models. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 339-358, Chapman and Hall.

[4] Berk, R. (1966). Limiting Behaviour of Posterior Distributions when the Model is Incorrect. *Annals of Mathematical Statistics 37*(1), 51–58.

[5] Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Chichester, UK: Wiley.

[6] Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics 7*, 686–690.

[7] Bingham, D. R. and H.A. Chipman (2003). Optimal designs for model selection. Technical Report, University of Michigan

[8] Borth, D. (1975). A total entropy criterion for the dual problem of model discrimination and parameter estimation. *Journal of the Royal Statistical Society, Series B 37*(1), 77–87.

[9] Box, G., and N. Draper (1987). *Empirical Model-Building and Response Surfaces*. New York: Wiley.

[10] Box, G. and W. Hill (1967). Discrimination among mechanistic models. *Technometrics 9*(1), 57–71.

[11] Box, G., W. Hunter, and J. Hunter (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building*. New York: Wiley.

[12] Box, G. and H.L Lucas (1959). Design of Experiments in Non-Linear Situations. *Biometrika 46*(1), 77–90.

[13] Cheng, R. C. H. and W. Holland (1997). Sensitivity of computer simulation experiments to errors in input data. *Journal on Statistical Compututing and Simulation 57,* 219–241.

[14] Chipman, H., M. Hamada. and C.F.J. Wu (1997). Bayesian Variable Selection for Designed Experiments with Complex Aliasing. *Technometrics 39*, 372–381.

[15] Chipman, H., E.I. Goerge. and R.E. McCulloch (2001). The Practical Implementation of Bayesian Model Selection. *IMS Lecture Notes - Monograph Series 38*, 67–116.

[16] DeGroot, M. (1962). Uncertainty, information and sequential experiments. *Annals of Statistics 33*, 404–419.

[17] Dellaportas, P. and A.F.M. Smith. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics 3*, 443–459.

[18] Dmochowski, J. (1996). Intrinsic Priors via Kullback-Leibler Geometry. *Bayesian Statistics 5*, 543–549.

[19] Fedorov, V. (1996). Discussion: Follow-up designs to resolve confounding in multifactor experiments (with discussion). *Technometrics 38*(4), 303–332.

[20] George, E. I. (1999). Bayesian model selection. *Encyclopedia of Statistical Sciences 3*, 39–46.

[21] Hill, P. (1978). A review of experimental design procedures for regression model discrimination. *Technometrics 20*(1), 15–21.

[22] Hill, W., W. Hunter, and D. Wichern (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics 10*(1), 145–160.

[23] Johnson, M. and C. Nachtsheim (1983). Some guidelines for constructing exact d-optimal designs on convex design spaces. *Technometrics 25*(3), 271–277.

[24] Kleijnen, J. (1996). Experimental design for sensitivity analysis, optimization, and validation of simulation models. In J. Banks (Ed.), *Handbook of Simulation*. New York: John Wiley & Sons, Inc.

[25] Law, A. M. and W. D. Kelton (2000). *Simulation Modeling & Analysis* (3d ed.). New York: McGraw-Hill, Inc.

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

[27] Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International Statistical Review 63*(2), 215–232.

[28] Meyer, R. D., D. M. Steinberg, and G. E. P. Box (1996). Follow-up designs to resolve confounding in multifactor experiments (with discussion). *Technometrics 38*(4), 303–332.

[29] Myers, R. H., R. H. Khuri, and W. H. C. Carter (1989). Response Surface Methodology: 1966-1988. *Technometrics 31*(2), 137–157.

[30] Myers, R. H., and D.C. Montgomery (1995). *Response Surface Methodology*. New York: Wiley.

[31] Ng, S.-H. (2001). Sensitivity and Uncertainty Analysis in Complex Simulation Models. *Ph.D. Thesis*. Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

[32] Ng, S.-H. and S. E. Chick (2001). Reducing input distribution uncertainty for simulations. In B. Peters, J. Smith, M. Rohrer, and D. Madeiros (Eds.), *Proceedings of the Winter Simulation Conference*, Piscataway, NJ, pp. in press. Institute of Electrical and Electronics Engineers, Inc.

[33] Nguyen, N. K., and A. J. Miller (1992). A Review of some exhange algorithms for constructing discrete D-optimal designs. *Computational Statistics and Data Analysis 14*, 489–498.

[34] Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

[35] Sanchez, S. M. (1994). A robust design tutorial. In J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila (Eds.), *Proceedings of the Winter Simulation Conference*, Piscataway, New Jersey, pp. 106–113. Institute of Electrical and Electronics Engineers, Inc.

[36] Schruben, L. W. and B. H. Margolin (1978). Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association 73*(363), 504–525.

[37] Shi, L., and S. Ólafsson. 2000. Nested partitions method for global optimization. *Operations Research* 48 (3): 424–435.

[38] Smith, A. F. M. and I. Verdinelli (1980). A note on Bayesian design for inference using hierarchical linear model. *Biometrika 67*(3), 613–619.

[39] Welch, W. J. (1982). Branch and bound search for experimental designs based on D-optimality and other criteria. *Technometrics 24*(1), 41–48.

[40] Wu, C.F. J., and M. Hamada (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley.