Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

JOINT CRITERION FOR FACTOR IDENTIFICATION AND PARAMETER ESTIMATION

Stephen E. Chick

Technology Management Area
INSEAD
77305 Fontainebleau Cedex, FRANCE

Szu Hui Ng

Singapore Institute of Manufacturing Technology
71 Nanyang Drive
Singapore 638075, SINGAPORE

ABSTRACT

One goal in simulation experimentation is to identify which input parameters most significantly influence the mean of simulation output. Another goal is to obtain good parameter estimates for a response model that quantifies how the mean output depends on influential input parameters. The majority of experimental design techniques focus on either one goal or the other. This paper uses a design criterion for follow-up experiments that jointly identifies the important parameters and reduces the variance of parameter estimates. The criterion is entropy-based, and is applied to a critical care facility simulation.

1 INTRODUCTION

Response metamodels that describe the relationship between inputs to a simulation and the output mean can be useful for design decisions based on simulation, and much focus has gone into selecting simulation inputs in a way that improves the estimate of the response model (Barton 1998; Kleijnen 1996; Law and Kelton 2000; Sanchez 1994). Response models can be used in iterative processes to identify design parameters (e.g., number of servers, line speeds) that optimize some expected reward criterion (e.g., mean monthly revenue). They can be useful to identify statistical input parameters whose uncertain values most greatly influence uncertainty about the mean system performance (Cheng and Holland 1997; Ng and Chick 2001; Jacobson, Buss, and Schruben 1991; Morrice and Schruben 1993).

There are a number of design criterion available to select input parameters for simulation experiments. Some, like the MD criterion of Box and Hill (1967), focus on designs that seek to identify the important input parameters. Others, like the D-optimal design criterion, determine designs that optimize the experiments ability to reduce uncertainty about estimates of parameters. Relatively little work has been done to develop experimental design criteria to both select input parameters to jointly identify important input parameters (also called *model selection*) and to improve parameter estimates. Exceptions include Borth (1975), who used a measurement of entropy, and Hill, Hunter, and Wichern (1968), whose criterion has several parameters to tune.

This paper describes a joint criterion for experimental design recently proposed in Ng and Chick (2002) that selects designs to simultaneously identify important parameters and to reduce the variance of those estimates. The joint criterion is based on entropy measurements, but overcomes an implicit problem in Borth's criterion regarding the calibration of entropy for model selection with respect to entropy for parameter estimation. The criterion simplifies to a closed form for the standard linear regression model with normal observation errors. Here we present the main ideas and preliminary experimental results. For a fuller treatment, see Ng (2001) or Ng and Chick (2002).

The criterion is most applicable to follow-up experiments after a preliminary set of runs has been completed. The criterion informs the selection of inputs for additional runs (design points) in a way that both resolves ambiguity about which inputs are most influential on the mean system output, and uncertainty in response parameter estimates.

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. Section 3 describes a system, and empirically evaluates the performance of the new joint design criterion with two other criteria. One describes model uncertainty alone, and the other tries to improve parameter uncertainty alone. The new criterion does well at both identifying important factors and reducing parameter uncertainty. The cost of computing the criterion, however, means that it is most useful in the context of simulations with particularly long run times.

2 FORMALISM

2.1 Design Space and Regression Models

The design criterion is applied here to a finite, but potentially very large, set of potential choices of inputs to a finite number n of simulation runs, where n is selected by the experimenter.

There are a finite number q of real-valued inputs, $x_1, \ldots, x_q$, each of which may be chosen to take on a finite set of different values. The inputs may be either design parameters, or parameters of input distributions (e.g. arrival rate, service time variance). The choice to focus on a finite set of values for each input requires that a discrete set of possible input values be chosen for inputs that may take on a continuum of values.

In order to allow for response models that are nonlinear in the inputs, the inputs 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 inputs. For example, if a linear response model is desired, then set $q = p$ and $y_i = x_i$ for $i = 1, \ldots, q$. For quadratic response models, set $p = q + q(q+1)/2$ to allow for additional predictors y_k involving cross terms $x_i x_j$ for $i \le j$.

We follow the formulation of Raftery, Madigan, and Hoeting (1997) 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. More formally, the output z_i of the i-th simulation run is presumed to be of the form

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

where $y_{(1)}, \ldots, y_{(t)}$ are t predictors selected from the set of p predictors, and ζ_i is a zero mean noise term. The selection of a candidate response model is intended to identify the important predictors in the model, relative to the size of the noise in the simulation output.

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 a $n \times q$ matrix whose i-th row contains the values of the inputs for the i-th run. For a model with t predictors, the design matrix can be converted to a $n \times t$ predictor matrix $\mathbf{y} = \mathbf{y}(\mathbf{x})$ whose rows contain the values of predictors for each run. The values of the predictors take on a finite set of values, as determined by the finite set of values that can be assumed by the inputs. 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 both identifying the most important predictors (i.e., selects the model in $\mathbf{M}$ with the most appropriate functional form), and to estimate its regression parameters.

To formalize this, we first describe uncertainty about both response model selection and parameter estimation with probability distributions, then select the design matrix that best improves an entropy-based design criterion.

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.

We make a standard assumption of jointly independent, normally distributed errors, $\zeta_i \sim \text{Normal}\left(0, \sigma^2\right)$. This means that if model i is the true model, $\boldsymbol{\beta}_i$ is the true parameter, and $\mathbf{y}$ is the predictor matrix of the design, then the output $\mathbf{Z}$ has an multivariate normal distribution, $\mathbf{Z} \sim \text{Normal}_n\left(\mathbf{y}\boldsymbol{\beta}_i, \sigma^2 I_n\right)$, where I_n is the identity matrix. Batching may be indicated if the simulation output is not roughly normally distributed.

For prior distributions, the analysis presumes the conjugate prior distribution (Bernardo and Smith 1994) for the unknown parameter $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2)$, conditional on the ℓ-th model M_ℓ being correct,

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

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 $\beta/(\alpha - 1)$.

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. 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 the matrix $\mathbf{V}_\ell$ 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 $(\nu\lambda + \left(\mathbf{z}_0 - \mathbf{y}_0\boldsymbol{\mu}'_\ell\right)^{\mathrm{T}} \mathbf{z}_0 + \left(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}'_\ell\right)^{\mathrm{T}} \mathbf{V}_\ell^{-1}\boldsymbol{\mu}_\ell)/2$.

The prior distribution for M_ℓ can be chosen to be discrete uniform $(p(M_\ell) = 1/s)$ to be noninformative. Closed form

formulas to update probabilities $p(M_\ell \mid \mathbf{z}_0, \mathbf{y}_0)$ based on preliminary runs are provided by Raftery, Madigan, and Hoeting (1997).

In the rest of the paper, we presume the distribution in Eq. (2) and the prior $p(M_\ell)$ for use in a follow-up stage of experiments are based on noninformative priors in combination with an initial stage.

2.2.2 Entropy-Based Criterion

Several authors (Lindley 1956; DeGroot 1962; Bernardo 1979) proposed the use of the expected gain in Shannon information (or decrease in entropy) given by an experiment as a optimal design criterion to select the values for the design factors for the experiment. The choice of design influences the expected gain in information, for example, because the predictive distribution of future output $\mathbf{Z}$ is determined by the predictor matrix $\mathbf{y}$ and the prior distribution in Eq. (2).

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

The marginal distribution of $\mathbf{Z}$ given $M_\ell, \mathbf{y}$ is obtained by integrating out over σ^2, and results in 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 (1967) proposed the use of 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.

$$J = \sum_{\ell=1}^{s} p(M_\ell) \times \int \log \frac{p(\mathbf{Z} \mid M_\ell, \mathbf{y})}{\sum_{l=1}^{s} p(\mathbf{Z} \mid M_l, \mathbf{y}) p(M_l)} p(\mathbf{Z} \mid M_\ell, \mathbf{y}) d\mathbf{Z}. \quad (3)$$

In general, a closed form is unknown, so Box and Hill (1967) gave an upper bound approximation for J, the expected gain in Shannon information between the predictive distributions of each pair of candidate models M_i and M_l. This approximation to J was originally named the D-criterion (Box and Hill 1967), but we use the notation MD, as does Meyer, Steinberg, and Box (1996).

$$MD = \sum_{0 \le i \ne l \le s} p(M_i) p(M_l) \times \left(\int \log \frac{p(\mathbf{Z} \mid M_i, \mathbf{y})}{p(\mathbf{Z} \mid M_l, \mathbf{y})} p(\mathbf{Z} \mid M_i, \mathbf{y}) d\mathbf{Z} \right). \quad (4)$$

The MD criterion has proved to be effective in practice and popular with research workers (Hill 1978).

We will use MD for the model discrimination portion of our joint criterion. For the normal linear model, Hill and Hunter (1969) and Meyer, Steinberg, and Box (1996) showed 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. It can be shown that a closed form also results if the conjugate prior is assumed.

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

$$\sum_{0 \le i \ne l \le s} \frac{1}{2} p(M_i) p(M_l) \left[-n + \text{tr}(\mathbf{V}_l^{*-1}\mathbf{V}_i^*) + \frac{1}{\lambda} (\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_l)^{\mathrm{T}} \mathbf{V}_l^{*-1} (\widehat{\mathbf{z}}_i - \widehat{\mathbf{z}}_l) \right]. \quad (5)$$

Proof. See Ng and Chick (2002).

For parameter estimation, Bernardo (1979) and Smith and Verdinelli (1980) 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.

$$BD = \int \int p(\mathbf{Z}) p(\boldsymbol{\theta} \mid \mathbf{Z}) \log \left[\frac{p(\boldsymbol{\theta} \mid \mathbf{Z})}{p(\boldsymbol{\theta})} \right] d\boldsymbol{\theta} d\mathbf{Z} \quad (6)$$

Eq. (6) 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. *F*or a linear model M_ℓ of the form Eq. (1), the prior probability model Eq. (2), and a given design $\mathbf{y}$,

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

Proof. See Ng and Chick (2002).

A measure S_P of the overall expected gain in information for parameter uncertainty averages the entropy from each candidate regression model. For the linear model, this is

$$S_P = \sum_{\ell=1}^{s} \frac{p(M_\ell)}{2} \log \left| \mathbf{y}^{\mathrm{T}}\mathbf{y} + \mathbf{V}_\ell^{-1} \right| - C \quad (7)$$

for some C that does not depend on the design.

In essence, Borth (1975) simply added entropy measures J and S_P to obtain a joint criterion. But this ignores the fact that entropy for continuous variables (parameter estimation) might not be directly commensurate with entropy for discrete variables (model selection). These measures may have

different ranges over the design space. We therefore are led to renormalize the two entropy measures before summing. And since it is hard to evaluate the J criterion, we approximate it with the MD criterion. Thus, an upper bound approximation of the joint criterion for model discrimination and parameter estimation is

$$S_Q = \frac{MD - MD_{min}}{MD_{max} - MD_{min}} + \frac{S_P - S_{P_{min}}}{S_{P_{max}} - S_{P_{min}}}, \qquad (8)$$

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 $\mathbf{D}$

To achieve the dual objective of model discrimination and parameter estimation, we seek to obtain a design $d \in \mathbf{D}$ that maximizes S_Q in Eq. (8). Eq. (8) equally weights the model and parameter criteria by normalizing over the possible range of values, max minus min, over all designs.

For normal linear models, S_Q simplifies to a closed form (Eq. (5), Eq. (7)).

2.3 Computational Issues

Although S_Q simplifies to a closed form for the normal linear model outlined above, there are three computational challenges. 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 select from.

To address the first challenge, the summands for S_P and MD are computed by using only the most likely models after the first stage. There are typically far fewer than $s = 2^p$ different models whose probability $p(M_\ell)$ will lead it to be a competitor for the 'best' after the initial stage of experimentation. When direct enumeration is not computationally feasible, these 'more important' models can be identified heuristically by using Markov chain monte Carlo model composition, (MC^3) (Madigan and York 1995) 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, Madigan, and Hoeting (1997). 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.

Second, we estimate $MD_{min}, MD_{max}, S_{P_{min}}, S_{P_{max}}$ by randomly sampling many designs, then setting the estimates to be the minimum and maximum entropy values sampled from among those designs.

Third, we use an optimization heuristic to identify a design with a high value of S_Q. Simulated annealing (Aarts and Korst 1989), nested partitions (Shi and Ólafsson 2000), and a number of other tools are available to help with this. Here, we generalize a version of the k-exchange algorithm of Johnson and Nachtsheim (1983). 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. The algorithm is essentially a greedy algorithm that swaps in and out design points one at a time. Although global optimality is not guaranteed, we restart the algorithm multiple times with random initial conditions to avoid local optima.

3 NUMERICAL RESULTS

3.1 Example: Critical Care Facility

The critical care facility illustrated in Figure 1 was originally studied by Schruben and Margolin (1978). Patients arrive according to a Poisson process and are routed through the system depending upon their specific health condition. Stays in the intensive care (ICU), coronary care (CCU), and intermediate care facilities are presumed to be lognormally distributed.

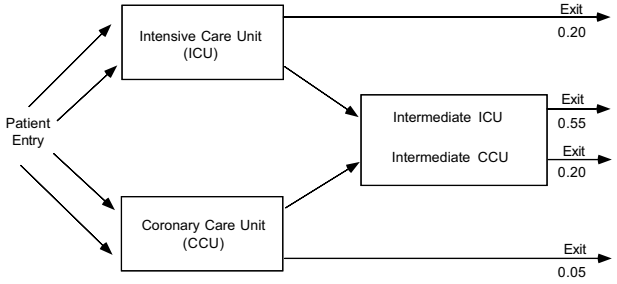

Figure 1: Fraction of Patients Routed through Different Units of a Critical Care Facility

Schruben and Margolin (1978) 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 $K = 1 + 4 * 2 + 3 = 12$ dimensions of parameters. The analysis here presumes a linear response model in these 12 parameters.

For the lognormal service times, the log has mean μ and precision $\lambda = 1/\sigma^2$. Subscripts distinguish the parameters of one service provider from each other (e.g., $\mu_{\text{icu}}, \mu_{\text{iicu}}, \mu_{\text{iccu}}, \mu_{\text{ccu}}, \lambda_{\text{icu}}, \ldots$).

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 Chick (2001), 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.

3.2 Results: Critical Care Facility

We initially ran a 32 run design with 2 replications for each of the 32 design points. We first analyzed the data corresponding to the 32 runs with the Bayesian analysis technique. We considered twelve input parameters, and thus 2^{12} distinct linear models in the model space **M**, each differing by the absence and presence of each predictor. Table 1 shows the posterior probabilities for the top 10 models. The model identified in a 512 run study in Ng and Chick (2001) is ranked 10th here.

Table 1: The Ten Most Probable Models after 32 Runs

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.513
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_3, p_4$	0.0936
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_4$	0.0539
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, p_1, p_4$	0.0427
$\lambda_{\text{sys}}, \lambda_{\text{ccu}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.0370
$\lambda_{\text{sys}}, \mu_{\text{ccu}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.0361
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \lambda_{\text{iccu}}, p_1, p_4$	0.0294
$\lambda_{\text{sys}}, \mu_{\text{icu}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.0275
$\lambda_{\text{sys}}, \lambda_{\text{icu}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.0262
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.0237

In order to determine a 48 run follow-up experiment to better discriminate among the top ten models and obtain more precise estimates for the model, we used the new S_Q criterion to determine the design. We used the design points of a full factorial for the 12 parameters as candidates for the 48 run design. As the number of models is large and the posterior probabilities of most of the models are very small, we renormalized the posterior probabilities of the top ten models, and considered only these models in the summation of S_Q. Since the number of possible 48 run designs from the 2^{12} candidate runs is large, we used k-exchange to determine the 'best' S_Q design. We used 50 randomly sampled designs to attempt to avoid local minima (and we used parameter $k = 5$ for the k-exchange algorithm).

Using the best design found with the k-exchange algorithm, we ran the critical care simulation again with 2 replications for each run. The posterior probabilities for the top three models, given the data from the combined design (32*2+48*2), are shown in Table 2. The top model is the same model identified in the 512 runs analysis in (Ng and Chick 2001), but the S_Q criterion enabled this model to be identified with far fewer runs.

Table 2: The Three Most Probable Models with 32 + 48 Runs Determined by the New Joint Criterion, S_Q

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.424
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.141
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_4$	0.140

The joint criterion S_Q results in different designs than the model discrimination criterion MD (Eq. (5)) and the parameter estimation criterion S_P (Eq. (7)). We use the k-exchange algorithm with similar settings (50 randomly sampled test designs, $k = 5$) to determine a good MD and S_P design. The posterior probabilities that result after running the MD design are presented in Table 3. The MD design identified the same top model as the S_Q design.

Table 3: The Three Most Probable Models with 32 + 48 Runs Determined by the MD Criterion

Model	Post. Prob.
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_3, p_4$	0.442
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, p_1, p_4$	0.118
$\lambda_{\text{sys}}, \mu_{\text{iicu}}, \lambda_{\text{iicu}}, \mu_{\text{iccu}}, p_1, p_4$	0.102

The MD criterion does a slightly better job in discriminating between the top two models than the S_Q design. However, Table 4 indicates that design S_Q did a better job than MD at reducing the parameter generalized variance (the determinant of the posterior covariance matrix of the parameter estimates, $|V(\boldsymbol{\beta})|$) of the top model.

Table 4: Comparison of S_Q and MD Criteria for A Posteriori Top Model after 32 + 48 Runs

Criterion	$\lvert V(\boldsymbol{\beta})\rvert$
S_Q	0.88×10^{-15}
MD	1.46×10^{-15}

The top two models identified in the S_P design are the same models identified in the original 32 run analysis. The posterior probabilities for these models are about the same

as the original analysis. The top model identified by the S_Q and MD design is only ranked eighth when the S_P design is used. As the top models differ, direct comparison of the generalized variance for the model cannot be made.

3.3 Discussion: Critical Care Facility

The results for this experiment illustrate the compromise between model discrimination and parameter estimation obtained when using the joint criterion S_Q. The S_Q design performed slightly poorer than the MD design for model discrimination, but was better in parameter estimation than MD. The S_P criteria focuses solely on the parameter estimates of the top models. As the posterior probability of the top model for S_Q (λ_{sys}, μ_{iicu}, λ_{iicu}, μ_{iccu}, p_1, p_3, p_4) was quite small after only 32 runs, the S_P criteria focused on designs that had good parameter estimation for only the top models with higher posterior probability. This example also shows that when there are quite a few input parameters and the output response is uncertain, with only a small number of runs (32 runs), using the S_P criterion may prematurely focus the attention on the top few models identified with this small number of runs for parameter estimation.

4 CONCLUSIONS

In this numerical example, the S_Q criterion did almost as well as the MD criterion for model discrimination, and did better in parameter estimation than MD. The top model for the S_P design was the same model found in the initial 32 runs (and different from the top model of S_Q and MD). This illustrates that using the S_P criterion too early in the experimentation process might prematurely focus the design and experimentation on one or two top models that may or may not be good approximations to the system (because of the small number of runs). Results from another application (data not shown) support this assessment.

More generally, the large number of matrix calculations required to compute the S_Q criterion needs to be balanced against the cost of actually running simulation replications. In the critical care facility simulation, the simulation itself requires a rather small amount of CPU time. CPU cycles might better be spent running replications rather than computing S_Q if the simulations are rather small. For complex simulations with long run times, however, the S_Q criterion may be an effective mechanism to balance the needs of factor identification and parameter estimation.

Future work includes allowing the weight between model and parameter uncertainty to vary through time in a sequential experimentation process, and the use of alternate discrete optimization tools for selecting the best design.

REFERENCES

Aarts, E., and J. Korst. 1989. *Simulated annealing and Boltzmann machines*. New York: John Wiley.

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

Bernardo, J. 1979. Expected information as expected utility. *Annals of Statistics* 7:686–690.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory*. Chichester, UK: Wiley.

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.

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

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.

DeGroot, M. 1962. Uncertainty, information and sequential experiments. *Annals of Statistics* 33:404–419.

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

Hill, W., and W. Hunter. 1969. A note on designs for model discrimination: Variance unknown case. *Technometrics* 11 (1): 396–400.

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.

Jacobson, S. H., A. H. Buss, and L. W. Schruben. 1991. Driving frequency selection for frequency domain simulation experiments. *Operations Research* 39:917–924.

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

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

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling & analysis*. 3d ed. New York: McGraw-Hill, Inc.

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

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

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.

Morrice, D. J., and L. W. Schruben. 1993. Simulation factor screening using harmonic analysis. *Management Science* 39:1459–1476.

Ng, S. H. 2001. *Sensitivity and uncertainty analysis of complex simulation models*. Ph. D. thesis, The University of Michigan, Ann Arbor, MI. Dept. of Industrial and Operations Engineering.

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

Ng, S.-H., and S. E. Chick. 2002. A Bayesian follow up design criterion for the dual objective of model discrimination and parameter estimation. Technical report, INSEAD.

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.

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

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.

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

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

AUTHOR BIOGRAPHIES

STEPHEN E. CHICK is Associate Professor of Technology Management at INSEAD in Fontainebleau, France. He is on leave from the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. In addition to stochastic simulation, his research focuses on Bayesian statistics, public health, decision analysis and computational methods in statistics. His research is motivated by problems in manufacturing, operations, and health care. He can be contacted at <stephen.chick@insead.edu>.

SZU HUI NG received her Ph.D. from the Industrial and Operations Engineering Department at the University of Michigan in 2001. Her research interests include stochastic simulation, design of experiments and Bayesian computation. She is currently a research fellow at the Singapore Institute of Manufacturing Technology in Singapore. Her email address is <ngsh@SIMTech.a-star.edu.sg>.