KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0326

USING THE BAYESIAN INFORMATION CRITERION TO DEVELOP TWO-STAGE MODEL-ROBUST AND MODEL-SENSITIVE DESIGNS

by
A. RUGGOO
M. VANDEBROEK

D/2003/2376/26

Using the Bayesian Information Criterion to develop two-stage model-robust and model-sensitive designs

Arvind Ruggoo
Martina Vandebroek

Department of Applied Economic Sciences, Katholieke Universiteit Leuven

Abstract

In this paper, we investigate use of the Bayesian Information Criterion (BIC) in the development of Bayesian two-stage designs robust to model uncertainty. The BIC is particularly appealing in this situation as it avoids the necessity of prior specification on the model parameters and can readily be computed from the output of standard statistical software packages.

Keywords: Two-stage procedures, BIC, prior probabilities, integrated likelihood, posterior probabilities, bias, lack of fit

1 Introduction

D-optimality (and alphabetic optimality criteria in general) has been criticized for being too dependent on the assumed model and for making no provision for model checking. Research in the recent years has concentrated on developing algorithms that retain the flexibility of the D-optimal approach but also reduce model dependence by providing protection against the bias induced by incorrect model specification and also making provision for detection of lack of fit. In that context, a recent development in the area is the two-stage procedure of Ruggoo and Vandebroek (2003), henceforth referred to as RUVA. They assume that the true model comprises some primary terms that will eventually be fitted, and some potential terms. In the first stage they use a criterion that facilitates the improvement of the proposed model by detecting lack of fit. The design in the second stage then uses model information from the first stage and attempts

to minimize bias with respect to potential terms. Their two-stage procedure generates designs with significantly smaller bias errors compared to standard single stage designs used in the literature. They also improve the coverage over the factor space. We now outline the development of the two-stage approach of RUVA in Section 2.

2 RUVA's two-stage design robust to model uncertainty

Suppose the linear model that will be fitted by the experimenter is of the form

$$y = \mathbf{x}'_{pri}\boldsymbol{\beta}_{pri} + \varepsilon$$

with $\mathbf{x}_{pri}$ being a p-dimensional vector of powers and products of the factors and $\boldsymbol{\beta}_{pri}$ the p-dimensional vector of unknown parameters attached to the primary terms. Let $\mathbf{x}'_{pot}\boldsymbol{\beta}_{pot}$ contain the terms that one wishes to protect against in designing the experiment, so that the model is actually of the form

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon = \mathbf{x}'_{pri}\boldsymbol{\beta}_{pri} + \mathbf{x}'_{pot}\boldsymbol{\beta}_{pot} + \varepsilon,$$

where $\mathbf{x}_{pot}$ is the q-dimensional vector containing powers and products of the factors not included in the fitted model and $\boldsymbol{\beta}_{pot}$ is the q-dimensional vector associated with the potential terms. The model is also reparametrized in terms of the orthonormal polynomials with respect to a measure μ on the design region. Since the primary terms are likely to be active and no particular directions of their effects are assumed, the coefficients of the primary terms are specified to have a diffuse prior distribution. On the other hand, potential terms are unlikely to have huge effects and the assumption $\boldsymbol{\beta}_{pot} \sim N(\mathbf{0}, \tau^2\sigma^2\mathbf{I}_q)$ proposed by DuMouchel and Jones (1994) (DMJ) is appropriate. The parameter τ^2 is the common prior variance of the potential terms' coefficients, measured in units of the random error variance σ^2. Following the orthonormalization procedure, which ensures that the effects are well separable and independent, the joint prior distribution assigned to $\boldsymbol{\beta}_{pri}$ and $\boldsymbol{\beta}_{pot}$ is $N(\mathbf{0}, \sigma^2\tau^2\mathbf{K}^{-1})$ where $\mathbf{K}$ is a $(p+q) \times (p+q)$ diagonal matrix, the first p diagonal elements of which are equal to zero and the remaining q diagonal elements are equal to one. Assume that $\mathbf{y}_i|\boldsymbol{\beta} \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I}_{n_i})$ for each stage i $(i = 1, 2)$ and that the first and second stage comprises n_1 and n_2 runs respectively so that the

total number of design points in the combined design is $n = n_1 + n_2$. $\mathbf{X}$ is the extended design matrix of dimension $n \times (p+q)$ for the combined stages, so that $\mathbf{X}' = [\ \mathbf{X}_1'\ \mathbf{X}_2'\ ]$. $\mathbf{X}_1 = [\ \mathbf{X}_{pri(1)}\ \mathbf{X}_{pot(1)}\ ]$ is of dimension $n_1 \times (p+q)$ and $\mathbf{X}_2 = [\ \mathbf{X}_{pri(2)}\ \mathbf{X}_{pot(2)}\ ]$ is of dimension $n_2 \times (p+q)$. They represent respectively the first and second stage designs expanded to full model space. $\mathbf{X}_{pri(i)}$ and $\mathbf{X}_{pot(i)}$ correspond to the primary and potential terms respectively for each stage i $(i = 1, 2)$.

Before observing the first stage data, the experimenter has specified a set of $(p+q)$ regressors defining the full model. The true relationship between the response and the input variables is believed to contain all primary terms and a subset q_i $(0 \leq q_i \leq q)$ of the potential terms. Consequently the total number of possible models is $m = 2^q$. Let us consider the subset models $M_1, M_2, \ldots, M_m$, with each model M_k defined by its corresponding parameters $\boldsymbol{\beta}_k$. RUVA assign prior probabilities, $p(M_i)$'s to each of the competing models using the effect inheritance assumption used in screening experiments, i.e. an interaction is more likely to be important if one or more of its parent factors is also important.

2.1 Development of the first stage design

A Bayesian first stage Generalized D (GD) optimal design for model M_k is the set of design points $\mathbf{X}_1^{(k)} = \left[\ \mathbf{X}_{pri(1)}^{(k)}\ \ \mathbf{X}_{pot(1)}^{(k)}\ \right]$ which minimizes

$$\mathrm{GD}_1^{(k)} = \left[\ \frac{1}{p}\log\left|\left(\mathbf{X}_{pri(1)}^{(k)\prime}\ \mathbf{X}_{pri(1)}^{(k)}\right)^{-1}\right| + \frac{\alpha_L}{q}\log\left|\left(\mathbf{L}_1^{(k)} + \frac{\mathbf{I}_q^{(k)}}{\tau^2}\right)^{-1}\right|\right] \tag{1}$$

(See RUVA for more details). It can be seen that criterion (1) is made up of two components; the first corresponding to precision of primary terms and the second has a weight α_L, to attach importance on the lack of fit expression to improve knowledge on the true model. $\mathbf{X}_{pri(1)}^{(k)}$, $\mathbf{L}_1^{(k)}$ and $\mathbf{I}_q^{(k)}$ are the matrices corresponding to $\mathbf{X}_{pri(1)}$, $\mathbf{L}_1$ and $\mathbf{I}_q$ expanded to model space $\mathbf{M}_k$ where

$$\mathbf{L}_1 = \mathbf{X}_{pot(1)}'\mathbf{X}_{pot(1)} - \mathbf{X}_{pot(1)}'\mathbf{X}_{pri(1)}\left(\mathbf{X}_{pri(1)}'\mathbf{X}_{pri(1)}\right)^{-1}\mathbf{X}_{pri(1)}'\mathbf{X}_{pot(1)},$$

is the familiar dispersion matrix encountered in the literature on model-sensitive designs.

It is interesting to note that the first stage design criterion is similar to the weighted combination of the D-optimum design for $\boldsymbol{\beta}_{pri}$ and the D_s-optimum design for $\boldsymbol{\beta}_{pot}$ suggested by Atkinson and Donev (1992). For some weight α $(0 \leq \alpha \leq 1)$, they propose to find exact designs for model M_k by minimizing

$$\frac{\alpha}{p} \log \left| \left(\mathbf{X}_{pri(1)}^{(k)\prime} \mathbf{X}_{pri(1)}^{(k)} \right)^{-1} \right| + \frac{1-\alpha}{q} \log \left| \mathbf{L}_1^{(k)} \right|^{-1}. \tag{2}$$

Their dual-purpose criterion ensures efficient estimation of parameters of the assumed primary model and detection of departures from that model. It is crucial to recognize that Atkinson and Donev (1992) composite design criterion does not depend on any form of prior assumption on the model parameters, $\boldsymbol{\beta}$. By adding the matrix $\frac{\mathbf{I}_q^{(k)}}{\tau^2}$ to $\mathbf{L}_1^{(k)}$ in (1), RUVA use the idea of DuMouchel and Jones (1994) to allow smaller design matrices and avoid singularity problems.

Since the prior probabilities, $p(M_i)$'s, reflect a priori model importance, RUVA incorporate them as weights in the first stage criterion so that the first stage design $\mathbf{X}_1 = [\, \mathbf{X}_{pri(1)} \, \mathbf{X}_{pot(1)} \,]$ is obtained by minimizing

$$\sum_{M_k} p(M_k) \, \mathrm{GD}_1^{(k)}.$$

2.2 Development of the second stage design

The Bayesian second stage GD optimal design for model M_k is the set of design points $\mathbf{X}_2^{(k)} = \left[\mathbf{X}_{pri(2)}^{(k)} \; \mathbf{X}_{pot(2)}^{(k)} \right]$ which minimizes

$$\mathrm{GD}_2^{(k)} = \left[\frac{1}{p} \log \left| \left(\mathbf{X}_{pri(1)}^{(k)\prime} \mathbf{X}_{pri(1)}^{(k)} + \mathbf{X}_{pri(2)}^{(k)\prime} \mathbf{X}_{pri(2)}^{(k)} \right)^{-1} \right| + \frac{\alpha_B}{q} \log \left| \mathbf{A}_2^{(k)\prime} \mathbf{A}_2^{(k)} + \mathbf{I}_q^{(k)} \right| \right], \tag{3}$$

where $\mathbf{X}_{pri(1)}^{(k)}$, $\mathbf{X}_{pri(2)}^{(k)}$, $\mathbf{A}_2^{(k)}$ and $\mathbf{I}_q^{(k)}$ are the matrices corresponding to $\mathbf{X}_{pri(1)}$, $\mathbf{X}_{pri(2)}$, $\mathbf{A}_2$ and $\mathbf{I}_q$ expanded to model space $\mathbf{M}_k$ and

$$\mathbf{A}_2 = \left(\mathbf{X}'_{pri(1)} \mathbf{X}_{pri(1)} + \mathbf{X}'_{pri(2)} \mathbf{X}_{pri(2)} \right)^{-1} \left(\mathbf{X}'_{pri(1)} \mathbf{X}_{pot(1)} + \mathbf{X}'_{pri(2)} \mathbf{X}_{pot(2)} \right)$$

is the alias matrix in the combined stage. The objective of the second stage is to use model information from first stage data to minimize bias with respect to potential terms.

Box and Meyer (1993) propose a general way for calculating the posterior probabilities of different candidate models within the framework of fractionated screening experiments. Given the first stage data $\mathbf{y}_1$, the posterior probability of model M_i given $\mathbf{y}_1$ is

$$p(M_i|\mathbf{y}_1) \propto p(M_i)p(\mathbf{y}_1|M_i), \tag{4}$$

where $p(M_i)$ is the prior probability of model M_i and $p(\mathbf{y}_1|M_i)$ is the integrated likelihood of $\mathbf{y}_1$ given model M_i. The resulting posterior probability for model M_i given $\mathbf{y}_1$ can then be obtained along the lines shown in Box and Meyer (1993):

$$p(M_i|\mathbf{y}_1) = C\, p(M_i)\, \tau^{-q_i} \left| \mathbf{X}_i'\mathbf{X}_i + \frac{\mathbf{K}_i}{\tau^2} \right|^{-1/2} \left(\mathbf{S}(\widehat{\boldsymbol{\beta}}_i) + \frac{1}{\tau^2}\, \widehat{\boldsymbol{\beta}}_i'\, \mathbf{K}_i\, \widehat{\boldsymbol{\beta}}_i \right)^{-(n_1-1)/2}, \tag{5}$$

where $\mathbf{X}_i$ is the first stage design in model M_i space and

$$\mathbf{K}_i = \begin{bmatrix} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times q_i} \\ \mathbf{0}_{q_i\times p} & \mathbf{I}_{q_i\times q_i} \end{bmatrix},$$

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i + \frac{\mathbf{K}_i}{\tau^2} \right)^{-1} \mathbf{X}_i'\mathbf{y}_1 = E(\boldsymbol{\beta}_i|\mathbf{y}_1), \text{assuming model } M_i,$$

$$\mathbf{S}(\widehat{\boldsymbol{\beta}}_i) = (\mathbf{y}_1 - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)'(\mathbf{y}_1 - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i) = \text{Residual Sum of Squares for model } M_i$$

and finally C is the normalization constant that forces all probabilities to sum to one.

Since the Box and Meyer posterior probabilities computed from first stage data in (5) reflect a posteriori model importance, RUVA incorporate them as weights to average the GD criterion in (3) when the second stage is selected. This is achieved by choosing the second stage design points $\mathbf{X}_2$ so as to minimize

$$\sum_{M_k} p(M_k|\mathbf{y}_1)\ \mathrm{GD}_2^{(k)}.$$

RUVA refers to this two-stage approach as the Bayesian MGD-MGD two-stage procedure, the acronym MGD enforcing the analogy that model uncertainty is taken care in the GD criterion in both stages by sweeping over the different possible models. RUVA's two-stage designs have good properties with respect to precision of important terms, lack of fit and bias properties with respect to a true assumed model in various simulation studies.

The two-stage procedure of RUVA is in-built within the realm of the Bayesian paradigm and consequently requires prior densities for the model parameters and also the prior model probabilities. As is often the case the problem of determining a prior distribution from available information is the most delicate matter in Bayesian methodology. Recall that the joint prior distribution assigned to $\boldsymbol{\beta}_{pri}$ and $\boldsymbol{\beta}_{pot}$ was $N(\mathbf{0}, \sigma^2\tau^2\mathbf{K}^{-1})$. Consequently the second stage procedure depends on the parameter τ^2, which controls both the individual integrated likelihoods, $p(\mathbf{y}_1|M_i)$'s and the adaptivity of the Box and Meyer posterior model probabilities in (5). Improper specification of τ^2 will affect the posterior weights used as measures of fit in the second stage criterion. RUVA propose to use the default value of $\tau = 1$ in both stages of the MGD-MGD approach to achieve satisfactory designs with respect to a combined criterion involving precision, lack of fit and bias properties.

The objective of this paper is to modify the second stage procedure of RUVA, so that it is independent on prior specification of the parameter τ^2. This will have the signal advantage of one less parameter to specify, which is usually hard to know a priori, when obtaining the two-stage designs. We shall for that purpose approximate the integrated likelihood, $p(\mathbf{y}_1|M_i)$, using the Bayesian Information Criterion (BIC). The BIC avoids the necessity of prior specification on the model parameters and is reasonable for many practical purposes. Consequently we shall compute a new set of posterior probabilities, $p(M_i|\mathbf{y}_1)$, $(i = 1, 2, \ldots, m)$ to be used as measures of fit in the second stage criterion.

The paper will be organized as follows: In Section 3, we briefly review the Bayesian approach to model averaging and present the BIC as a simple and accurate approximation to the integrated likelihood. We recast the MGD-MGD procedure of RUVA so that it is independent of specification of the parameter τ^2 in Section 4. The performance of our modified two-stage procedure is evaluated in Section 5, and we show that it yields very good and comparable results to the MGD-MGD procedure of RUVA and also to the classical single stage D-optimal and Bayesian D-optimal procedures. We end with a conclusion in Section 6.

3 The Bayesian approach to model averaging

The appealing characteristic of the two-stage procedure of RUVA is the incorporation of model uncertainty by averaging the criterion over all possible models in both stages. In doing so, the procedure allows incorporation of several competing models and does not depend on specification of a single model. This approach can be thought of as a particular case of the Bayesian Model Averaging (BMA) procedures reviewed by Hoeting, Madigan, Raftery and Volinsky (1999). BMA provides a mechanism to account for model uncertainty by estimating some quantity under each model and then averaging the estimates according to how likely the model is (Wasserman, 1997). In the context of the two-stage procedures, the quantities are the posterior model probabilities computed from (5), that let the data give the competing models different weights of evidence, and are then used as measures of fit to average the GD-optimality criterion in the second stage. Madigan and Raftery (1994) note that averaging over all models in this fashion provides better average predictive ability than using any single model.

Let us consider the subset models $M_1, M_2, \ldots, M_m$ described previously with each model M_k defined by its corresponding parameters $\boldsymbol{\beta}_k$. Each model consists of a set of probability densities for the random variable $\mathbf{y}$. Once we obtain first stage data $\mathbf{y}_1$, the posterior probability for model M_j can be easily evaluated from Bayes' theorem and is

$$p(M_j|\mathbf{y}_1) = \frac{p(M_j)p(\mathbf{y}_1|M_j)}{\sum_r p(M_r)p(\mathbf{y}_1|M_r)}. \tag{6}$$

From classical probability theory, $p(\mathbf{y}_1|M_j)$ can be obtained by integrating over $\boldsymbol{\beta}_j$, i.e.

$$p(\mathbf{y}_1|M_j) = \int p(\mathbf{y}_1|\boldsymbol{\beta}_j, M_j)p(\boldsymbol{\beta}_j|M_j)\, d\boldsymbol{\beta}_j,$$

where $p(\mathbf{y}_1|\boldsymbol{\beta}_j, M_j)$ is the likelihood function for model M_j. The quantity $p(\mathbf{y}_1|M_j)$ is usually called the integrated likelihood for model M_j. Evaluating the integrated likelihood involves specifying priors for $\boldsymbol{\beta}_j$ and usually complex integration. The prior and integration problems can be solved in the following manner for regular statistical problems:

Let $\hat{\ell}_j = \ell_j(\widehat{\boldsymbol{\beta}}_j)$ denote the maximized log-likelihoods under model M_j and d_j be the dimension of $\boldsymbol{\beta}_j$. Then

$$\mathcal{S} = \hat{\ell}_j - \frac{d_j}{2}\log(n_1), \tag{7}$$

is a fairly accurate approximation of $\log p(\mathbf{y}_1|M_j)$ for a specific choice of prior called the "unit-information prior" on the parameter space, that says that the amount of information in the prior equals to the amount of information in one observation (See Kass and Wasserman, 1995, for more details). Raftery (1996) gives further evidence for the accuracy of this approximation. We observe that expression (7) is the familiar Schwarz criterion (Schwarz, 1978) and minus twice the Schwarz criterion is often referred to as the BIC (Kass and Raftery, 1995), i.e.

$$\mathrm{BIC_j} = -2\,\mathcal{S} \approx -2\log p(\mathbf{y}_1|M_j). \tag{8}$$

From (8), we have an approximate but easy way of obtaining the integrated likelihood which does not depend on prior specification of the model parameters and

$$p(\mathbf{y}_1|M_j) \approx e^{-0.5\,\mathrm{BIC_j}}. \tag{9}$$

For the linear regression with normal errors, Raftery (1995) shows that the most convenient form of BIC is

$$\mathrm{BIC}_j = n_1\log(1 - R_j^2) + k_j\log(n_1), \tag{10}$$

where R_j^2 is the usual R^2 (coefficient of determination) value for model M_j and k_j is the number of regressors (not including the intercept) in the model. Using our results from (9) and (10), the posterior probability of each model is easily found from (6) to be

$$p(M_j|\mathbf{y}_1) = C\,p(M_j)\,e^{-0.5\,\mathrm{BIC}_j}, \tag{11}$$

where C is a normalizing constant that forces all posterior probabilities to sum to unity.

The expression in (9) indicates that BIC is a Bayesian procedure that does not require the specification of a prior on model parameters, but provides a way to obtain an accurate approximation of the integrated likelihood. As explained by Kass and Wasserman

(1995), the BIC uses an implicit unit information prior, i.e. a multivariate normal prior with mean at the maximum likelihood estimate and the amount of information in the prior equal to the average amount of information in one observation. Since the prior is based on only one observation, it is vague yet proper. Further it involves readily available regression statistics for all candidate models which can be obtained from the output of standard statistical software packages. The appealing property of BIC which avoids prior specification on model parameters continues to be investigated and justified, see for example Pauler (1998) who motivates the BIC and propose two useful modifications of the criterion applicable to other types of problems. Recently Volinsky and Raftery (2000) have investigated BIC for variable selection in models for censored survival data.

Wasserman (1997) also shows that if model M_j denote the model containing the true density, i.e. the model that generates the data, then for $i \neq j$ and under weak conditions,

$$\frac{p(M_i|\mathbf{y}_1)}{p(M_j|\mathbf{y}_1)} \rightarrow 0$$

in probability. This means that the posterior probability of the true model goes to one and the posterior probabilities of the other models go to zero. Further Wasserman (1997) indicates that the BIC has the same asymptotic behavior as it selects the true model asymptotically. From this result, we would expect the BIC to provide a close approximation to the posterior probabilities and consequently satisfactory results in the second stage procedure.

4 Development of the two-stage procedures using the BIC

We now recast the two-stage MGD-MGD procedure of RUVA and use the posterior model formulation from (11) in our second stage procedure.

As argued in Section 2.1, we can view the first stage design as an extension of the composite design criterion of Atkinson and Donev (1992), which does not depend on specification of τ^2. We thus propose the first stage MGD optimal design $\mathbf{X}_1 = [\ \mathbf{X}_{pri(1)}\ \mathbf{X}_{pot(1)}\ ]$ to be

obtained by minimizing

$$\sum_{M_k} p(M_k)\ \mathrm{GD}_1^{(k)},$$

where

$$\mathrm{GD}_1^{(k)} = \left[\frac{1}{p} \log \left| \left(\mathbf{X}_{pri(1)}^{(k)'}\ \mathbf{X}_{pri(1)}^{(k)} \right)^{-1} \right| + \frac{\alpha_L}{q} \left| \mathbf{L}_1^{(k)} \right|^{-1} \right]. \tag{12}$$

$\mathbf{X}_{pri(1)}^{(k)}$ and $\mathbf{L}_1^{(k)}$ are the matrices corresponding to $\mathbf{X}_{pri(1)}$ and $\mathbf{L}_1$ expanded to model space $\mathbf{M}_k$. In practice, we observe that $\mathbf{L}_1^{(k)}$ may be singular. The problem is avoided in the algorithmic construction of designs by the addition of a small multiple of the identity matrix. That is, we let

$$\mathbf{L}_1^{(k)} = \mathbf{L}_1^{(k)} + \epsilon\, \mathbf{I}_q^{(k)},$$

where ϵ is a small number typically between 10^{-4} and 10^{-6}. Such type of matrix regularization is common in the construction of exact D-optimal designs (see for e.g. Atkinson and Donev (1992), Chapters 10 & 15).

The second stage design is then obtained by choosing the second stage design points $\mathbf{X}_2$ so as to minimize

$$\sum_{M_k} p(M_k|\mathbf{y}_1)\ \mathrm{GD}_2^{(k)},$$

where $p(M_k|\mathbf{y}_1)$ is computed for each model using (11) instead of (5) and $\mathrm{GD}_2^{(k)}$ is as in (3).

The modifications in the two-stage MGD-MGD procedure implicitly avoid specification of the additional parameter τ^2 in the second stage optimality criterion and provide, as we shall see later, an attractive alternative to the procedure of RUVA.

5 Evaluation of the two-stage procedures

The performance of our Bayesian two-stage procedure presented in Section 4, will now be evaluated relative to the two-stage approach of RUVA and the classical one-stage designs.

Since the second stage design is dependent on first stage data through the posterior model probabilities, a simulation approach is required. The performance of each design will be measured by its efficiency relative to a true assumed model in 200 simulations. The error $\varepsilon \sim N(0,1)$ is assumed in all the simulations. The unique stage competitors to the Bayesian two-stage optimal design are the traditional D-optimal design for the primary terms model and the Bayesian D-optimal design of DMJ.

As proposed by RUVA, the values of the following determinants will be used as measures of efficiency of the precision, lack of fit and bias components. The measure of precision of the primary terms is given by $D^*_{X_{pri}} = \left|\mathbf{X}^{*'}_{pri}\,\mathbf{X}^*_{pri}\right|^{-1/p}$, a measure of the lack of fit component is $D^*_{lof} = |\mathbf{L}^*|^{-1/q}$ and $D^*_{bias} = \left|\mathbf{A}^{*'}\mathbf{A}^* + \mathbf{I}_q\right|^{1/q}$ represents the degree of bias, where

$$\mathbf{L}^* = \mathbf{X}^{*'}_{pot}\mathbf{X}^*_{pot} - \mathbf{X}^{*'}_{pot}\mathbf{X}^*_{pri}\left(\mathbf{X}^{*'}_{pri}\mathbf{X}^*_{pri}\right)^{-1}\mathbf{X}^{*'}_{pri}\mathbf{X}_{pot} \quad \text{and} \quad \mathbf{A}^* = \left(\mathbf{X}^{*'}_{pri}\mathbf{X}^*_{pri}\right)^{-1}\mathbf{X}^{*'}_{pri}\mathbf{X}^*_{pot}.$$

$\mathbf{X}^*_{pri}$ and $\mathbf{X}^*_{pot}$ represent the combined first and second stage design points for the primary and potential terms expanded to contain regressors in the true model only. $D^*_{X_{pri}}$, D^*_{lof} and D^*_{bias} have been defined such that the smaller the value obtained, the better the design performs with respect to that criterion. The minimum bias design arises when $\mathbf{A}^* = \mathbf{0}$ and so $D^*_{bias} = |\mathbf{I}_q|^{1/q} = 1$.

The performance of the two-stage procedures are then measured by the average of $D^*_{X_{pri}}$, D^*_{lof} and D^*_{bias} over the 200 different simulations, i.e.

$$AD^*_{X_{pri}} = \frac{\sum_{i=1}^{200} D^*_{X_{pri}}}{200}, \qquad AD^*_{lof} = \frac{\sum_{i=1}^{200} D^*_{lof}}{200}, \qquad AD^*_{bias} = \frac{\sum_{i=1}^{200} D^*_{bias}}{200}.$$

The one-stage traditional non-Bayesian D-optimal design and one-stage Bayesian D-optimal design of DMJ are not data dependent and can thus be evaluated over the n design runs by the single measures $D^*_{X_{pri}}$, D^*_{lof} and D^*_{bias} for the true model. In connection with sample sizes for each stage, RUVA suggest using two-stage designs of size $n = 2(p+q+2)$ with half of the design points allocated to each stage of the design. $\alpha_L = 20$ is the default value used in the first stage and $\alpha_B = 10$ is used in the second stage. RUVA argues that these choices leads to satisfactory designs with respect to a

combined criterion involving precision, lack of fit and bias properties.

To enable comparison as to how our BIC based posterior model probabilities perform compared to the Box and Meyer probabilities used by RUVA in their two-stage approach, we shall consider first stage data simulated from the same true models in the three cases outlined by RUVA. The design region they consider is the $5 \times 5 \times 5$ grid on $[-1, +1]^3$.

Case I :

The true model from which first stage data is simulated is

$$y = 42.0 + 11.5\ x_1 + 12.8\ x_2 + 10.5\ x_3 + 14.6\ x_1^2 - 7.4\ x_2^2 + \varepsilon.$$

The true model comprises all the five primary terms, $\{1,\ x_1,\ x_2,\ x_3,\ x_1^2\}$ and one component, namely the quadratic effect of x_2, from the three potential terms, $\{x_1x_2,\ x_2^2,\ x_3^2\}$.

Case II :

RUVA consider in Case II a model with $p = 5$ primary terms, $\{1,\ x_1,\ x_2,\ x_3,\ x_1x_2\}$ and $q = 4$ potential terms, $\{x_1^2,\ x_1x_3,\ x_2^2,\ x_3^2\}$. First stage data is then simulated from

$$y = 42.0 + 11.2\ x_1 + 14.5\ x_2 + 10.6\ x_3 + 12.5\ x_1x_2 + 8.9\ x_1^2 - 9.9\ x_1x_3 + \varepsilon.$$

Case III :

Finally in this case, RUVA examine data simulated from

$$y = 40.0 + 11.5\ x_1 + 12.8\ x_2 + 10.5\ x_3 + 14.6\ x_1^2 + 9.8\ x_1x_2 - 7.4\ x_1x_3 - 8.7\ x_2^2 + \varepsilon$$

and with the full model comprising five primary terms namely, $\{1,\ x_1,\ x_2,\ x_3,\ x_1^2\}$ and an additional five potential terms, $\{x_1x_2,\ x_1x_3,\ x_2x_3,\ x_2^2,\ x_3^2\}$.

We present in Table 1, the prior and posterior model probabilities for the $m = 2^3 = 8$ possible models from the five primary and three potential terms for Case I. The last two columns of Table 1 correspond respectively to the posterior probabilities obtained using (11) and (5). The results shown are from one simulated first stage data set only. Other simulations showed similar good results. It is interesting to see that the BIC provides a

Table 1: BIC and Box & Meyer based posterior model probabilities

Plausible Models	Prior Probabilities	Posterior Probabilities	
		BIC	Box & Meyer
$1\ x_1\ x_2\ x_3\ x_1^2$ (Primary model)	0.5787037	0	0.0008668
$1\ x_1\ x_2\ x_3\ x_1^2\ x_1x_2$	0.1157407	0	0.0000427
$1\ x_1\ x_2\ x_3\ x_1^2\ x_2^2$ (True model)	0.1157407	0.7374309	0.8383601
$1\ x_1\ x_2\ x_3\ x_1^2\ x_3^2$	0.1157407	0	0.0010012
$1\ x_1\ x_2\ x_3\ x_1^2\ x_1x_2\ x_2^2$	0.0231481	0.1406869	0.0424615
$1\ x_1\ x_2\ x_3\ x_1^2\ x_1x_2\ x_3^2$	0.0231481	0	0.0000526
$1\ x_1\ x_2\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.0231481	0.0905888	0.1112121
$1\ x_1\ x_2\ x_3\ x_1^2\ x_1x_2\ x_2^2\ x_3^2$	0.0046296	0.0312934	0.006003

very good approximation to the integrated likelihood as reflected by the fact that the posterior probability of the true model is largest. In general the Box and Meyer probability is larger than the BIC based posterior probability for the true model. Intuitively, this is expected as the Box and Meyer posterior probabilities are more accurate since they involve actual integration of the integrated likelihood and also additional prior information on the model parameters in their computations.

The results of the evaluations for all the cases are shown in Tables 2 to 4. Using the BIC as an approximation to the marginal likelihood is reassuring as it gives very good and comparable results to the ones obtained by RUVA. The approach also gives excellent reductions in the bias in all three cases when compared to the unique stage D-optimal and Bayesian D-optimal design, whilst still maintaining good precision of the estimation of the effects of the primary terms. Interestingly the bias is smaller than the ones of RUVA in Cases II and III. The injection of additional prior information in the MGD-MGD procedure of RUVA may account for the slightly better precision for the effects of the primary terms in all the cases.

Table 2: Comparison of the two-stage procedure of RUVA and the one developed using BIC with the single stage design procedures.

Case I $y = 42.0 + 11.5\ x_1 + 12.8\ x_2 + 10.5\ x_3 + 14.6\ x_1^2 - 7.4\ x_2^2 + \varepsilon$.

Two-Stage Approach ($n_1 = n_2 = 10$)	$AD^*_{X_{pri}}$	AD^*_{lof}	AD^*_{bias}
MGD-MGD (RUVA)	0.046084	0.046428	1.004525
MGD-MGD (BIC)	0.048125	0.046865	1.004937
One-Stage Approach ($n = 20$)	$D^*_{X_{pri}}$	D^*_{lof}	D^*_{bias}
D-optimal (Primary Terms)	0.034299	-	2.428570
DuMouchel & Jones (1994)	0.038914	0.049374	1.279301

Table 3: Comparison of the two-stage procedure of RUVA and the one developed using BIC with the single stage design procedures.

Case II $y = 42.0 + 11.2\ x_1 + 14.5\ x_2 + 10.6\ x_3 + 12.5\ x_1x_2 + 8.9\ x_1^2 - 9.9\ x_1x_3 + \varepsilon$.

Two-Stage Approach ($n_1 = n_2 = 11$)	$AD^*_{X_{pri}}$	AD^*_{lof}	AD^*_{bias}
MGD-MGD (RUVA)	0.036782	0.036739	1.008554
MGD-MGD (BIC)	0.041773	0.039452	1.006493
One-Stage Approach ($n = 22$)	$D^*_{X_{pri}}$	D^*_{lof}	D^*_{bias}
D-optimal (Primary Terms)	0.022887	-	1.581590
DuMouchel & Jones (1994)	0.02958	0.031216	1.273629

Table 4: Comparison of the two-stage procedure of RUVA and the one developed using BIC with the single stage design procedures.

Case III $y = 40.0 + 11.5\ x_1 + 12.8\ x_2 + 10.5\ x_3 + 14.6\ x_1^2 + 9.8\ x_1x_2 - 7.4\ x_1x_3 - 8.7\ x_2^2 + \varepsilon.$

Two-Stage Approach ($n_1 = n_2 = 12$)	$AD^*_{X_{pri}}$	AD^*_{lof}	AD^*_{bias}
MGD-MGD (RUVA)	0.037010	0.031256	1.006440
MGD-MGD (BIC)	0.037740	0.033366	1.004922
One-Stage Approach ($n = 24$)	$D^*_{X_{pri}}$	D^*_{lof}	D^*_{bias}
D-optimal (Primary Terms)	0.028421	-	1.344158
DuMouchel & Jones (1994)	0.031606	0.023785	1.135410

6 Conclusions

We are all aware of the criticism of the dependence on an assumed model for the class of alphabetic optimal designs. Experimenters rarely have a model in hand and are faced with several competing candidate models. Clearly an algorithmic procedure that encompasses different possible model is desirable. The two-stage procedure we study, borrows tools from Bayesian methods and accounts for model uncertainty by considering all possible competing models. In this way we are not confined to defend any specific model in our criterion. We are currently unaware of any design procedure that explicitly uses the BIC to attack model uncertainty in experimental design problems in this way. The fact that computation of the BIC does not require introduction of prior distributions on the model parameters and relies on the unit-information which is the average amount of information in one observation, makes it intuitively appealing. We encourage further research on the BIC in work on optimal designs.

Acknowledgements

The authors wish to thank Dr. Peter Goos for some suggestions on an earlier draft of the article.

References

Atkinson, A. C. and Donev, A. N. (1992) Optimum Experimental Designs, Clarendon Press: Oxford.

Box, G. E. P. and Meyer, R. D. (1993) Finding the active factors in fractionated screening experiments, Journal of Quality Technology, vol 25, pp 94-105.

DuMouchel, W. and Jones, B. (1994) A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model, Technometrics, vol 36, pp 37-47.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial, Statistical Science, vol 14, pp 382-417.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors, Journal of the American Statistical Association, vol 90, pp 773-795.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion, Journal of the American Statistical Association, vol 90, pp 928-934.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty using Occam's window, Journal of the American Statistical Association, vol 89, pp 1535-1546.

Pauler, D. (1998) The Schwarz Criterion and related methods for the normal linear model, Biometrika, vol 85, pp 13-27.

Raftery, A. E. (1995) Bayesian model selection in social research. In Sociological Methodology, P. V. Mardsen (editor), vol 25, pp 111-195, Blackwells, Cambridge.

Raftery, A. E. (1996) Approximate Bayes factors for accounting for model uncertainty in generalised linear models, Biometrika, vol 83, pp 251-266.

Ruggoo, A. and Vandebroek, M. (2003) Two-stage designs robust to model uncertainty, Technical Report 0319, Department of Applied Economic Sciences, Katholieke Universiteit Leuven.

Schwarz, G. (1978) Estimating the dimension of a model, Annals of Statistics, vol 6, pp 461-464.

Volinsky, C. T. and Raftery, A. E. (2000) Bayesian information criterion for censored survival models, Biometrics, vol 56, pp 256-262.

Wasserman, L. (1997) Bayesian model selection and model averaging, Technical Report 666, Department of Statistics, Carnegie Mellon University.