Prediction-Based Model Selection for Bayesian Multiple Regression Models

Adam L. Pintar
Statistical Engineering Division
National Institute of Standards and Technology
Gaithersbug, MD 20899
(*adam.pintar@nist.gov*)

Christine M. Anderson-Cook
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545
(*c-and-cook@lanl.gov*)

Huaiqing Wu
Department of Statistics
Iowa State University
Ames, IA 50011
(*isuhwu@iastate.edu*)

Abstract

Model selection is an important part of model building for Bayesian linear models when the number of possible model terms is large. Most current approaches focus on posterior model probabilities or the deviance information criterion. This article proposes an alternative strategy that considers how the model will be used after its selection and selects models based on their predictive abilities over a user-specified portion of the covariate space defined by a joint probability distribution called the distribution of interest. Because it is difficult to summarize the "goodness" of a model with a single number, we present a suite of numerical and graphical tools for detailed comparisons of different models. These tools help select a best model or a collection of good models

based on their prediction performances over covariate locations likely to arise from the distribution of interest. The proposed method is illustrated with two examples. The first example motivates and illustrates the new method, while the second example considers what to do when comparing thousands of models. Simulation results demonstrate where the new method produces improvements in prediction ability over some existing methods.

Keywords: Bayesian Model Averaging, Correlated Variables, Deviance Information Criterion, Posterior Probability, Variable Selection

1 Introduction

Model selection is an important step in the process of building a Bayesian multiple regression model. If too many predictors are included, the spread of posterior distributions for model parameters or predictions of new observations may by unnecessarily large. If too few predictors are included, the posterior distributions may lead to biased point and interval estimates. This article considers a new model selection methodology for Bayesian multiple regression models with a focus on obtaining good prediction in a user-specified portion of the covariate space.

1.1 Motivation

We begin by defining the Bayesian multiple regression model (BMRM):

$$\begin{aligned} \left(y_i|\boldsymbol{\beta}^m, \sigma_m^2\right) &\sim N[(\boldsymbol{\beta}^m)'\boldsymbol{x}_i^m, \sigma_m^2],\ i = 1, 2, \cdots, n \\ \left(\boldsymbol{\beta}^m, \sigma_m^2\right) &\sim g^m, \end{aligned} \tag{1}$$

where $m = 1, 2, \cdots, N_{mod}$ indexes the models under consideration. Note that g^m serves generically as the joint prior probability density function (pdf) for $\boldsymbol{\beta}^m$ and σ_m^2. In Section 2, sensible forms for g^m are considered.

Several procedures for model selection exist for BMRM's, but all of those procedures consider only the *observed data*. If the primary goal of building

a BMRM is prediction at user-specified covariate locations, then this goal should be factored into the model selection process. In Section 2 we consider an example where the warehouse manager receiving shipments uses the number of drums and the total weight of the shipment to predict the time required to process the shipment. If the manager wishes to predict handling times for larger-than-typical shipments, the goal is to select a model that predicts well outside of the observed data range. Because the number of drums and the total weight of the shipment are correlated, careful thoughts are required for defining suitable (number of drums, total weight) pairs that are larger than typical. The ability of the new method, called the prediction-based model selection method (PBMSM), to handle such situations distinguishes it from currently available methods.

The above motivation is based on the need to extrapolate. The dangers in extrapolation are well known and documented by statisticians. However, in practice, extrapolation is sometimes necessary, and we are not promoting the use of statistical methods for extrapolating except when answering the key questions of interest requires it. Given that this is sometimes necessary, we think that this goal should be integrated into the model selection process. However, the PBMSM is not limited to extrapolation. Our second example considers good prediction in the design region from a designed experiment.

The PBMSM leverages the very powerful Bayesian model averaging (BMA) approach to prediction [21]. In fact, the PBMSM uses the predictions from BMA as the basis on which all models under consideration are compared. This leads one naturally to ask why not just use the predictions from BMA and stop. There are several reasons. First, with BMA the ability to interpret the regression parameters is lost. Second, a single model can often provide predictions that are close to the quality of the predictions from BMA with the added benefit of a much simpler model. Finally, once the selection phase is completed, computing predictions from a single regression model instead of an average of many is more straightforward.

1.2 Existing Work

We now review current model selection procedures for BMRM's, methods for quantifying the discrepancy between functions, and some graphical tools. Discrepancy measures and graphical tools relate to the new method because they form the basis of the model comparisons.

This review of model selection procedures for BMRM's is not meant to be exhaustive, but considers key selection procedures to which the PBMSM is compared. The deviance information criterion (DIC) [24] can be described as a measure of a model's fit to the data plus a penalty for model complexity. This interpretation is similar to the Akaike information criterion (AIC) [1] and Bayesian information criterion (BIC) [23] with the best models having lower values. The DIC draws a nice connection between frequentist and Bayesian model selection methods and. However, a more popular and intuitive procedure exists in the Bayesian paradigm.

In the Bayesian paradigm, it is natural to cast the model as another parameter and calculate its posterior probability. Once posterior model probabilities are calculated, several approaches exist. One approach is simply to select the model with the highest posterior probability. Another approach is to select the median posterior model (MPM) [3], which includes model terms with posterior probability greater than 0.5. Under certain conditions, in [3], it is argued that the MPM is optimal for prediction.

Another procedure in [12] assigns each regression coefficient a prior distribution that depends on model m. If the regression coefficient is absent from model m, its prior distribution is normal with mean zero and small variance. If the regression coefficient is present in model m, its prior distribution is normal with mean zero and larger variance. Then, g^m is taken as the product of the marginal prior distributions. The prior distribution for σ^2 does not change with m, but is chosen to have a convenient form.

Some other model selection algorithms aimed at prediction focus on the Bayesian posterior predictive distribution. Some examples are the predictive sample reuse technique of [9], the utility function approach of [22], the L, M, and K, criteria of [14] and minimizing expected posterior predictive loss as

described in [10].

1.3 Brief Description of the Algorithm

When using the Bayesian paradigm for inference, decisions about the predictive ability of a model are based on posterior distributions. Our goal is to identify a single model that precisely and accurately estimates a quantity of interest. Let Δ_{m_1} and Δ_{m_2} represent a common quantity of interest estimated from models m_1 and m_2, respectively. In the warehouse example, this is the mean time to process a shipment. Suppose Δ_0 is the true, but generally unknown, value of that quantity. Now, let F_{m_1} and F_{m_2} represent the posterior cumulative distribution functions (cdf's) for Δ_{m_1} and Δ_{m_2}, respectively, and let F be a step function that steps from 0 to 1 at Δ_0. Thus, F is a cdf or probability measure that places unit probability at Δ_0. Since we wish to select a single model to be used for prediction, the model we select for predicting Δ_0 is the one with the posterior cdf that most closely approximates F. To compare the relative performance of models, we propose a distance or discrepancy measure between the posterior cdf's and F related to the L_k distance between functions, see pages 90 and 91 of [2].

In the PBMSM, graphical tools are used to compare distributions of discrepancy measures. Boxplots are effective in making rough comparisons among distributions, and the fraction of covariate distribution (FCD) plots allow for finer distinctions among models. The FCD plot is similar to the fraction of design space (FDS) plot introduced by [25]. In [19], boxplots and FCD plots are used to examine the distributions of prediction mean squared error (MSE). In [18], FDS plots are used to compare predictions in the design space among competing designs for generalized linear models.

The remainder of the article is organized as follows: Section 2 describes the PBMSM with an example. Section 3 presents a simulation study, which compares the PBMSM to existing methods with respect to the prediction ability. Section 4 provides a second example with a large number of candidate models. Section 5 gives some concluding remarks.

2 Methodology

The new methodology, PBMSM, is described generally by a sequence of four steps:

1. Select and characterize the user-specified distribution of interest over the covariate space.

2. Sample points randomly from that distribution.

3. Estimate the discrepancy between the posterior distribution of the quantity being predicted and the ideal value at each point sampled in step 2 for each model under consideration.

4. Compare models graphically based on the discrepancy estimates to select a best model or a group of models.

These steps match those presented in [19] for the frequentist paradigm, with one important distinction. In the Bayesian paradigm we use the entire posterior distribution at each location to compare competing models instead of the mean squared error, which only focuses on variance and bias. A discrepancy measure is proposed between the posterior cdf of the quantity of interest for a particular model and the ideal value of that quantity. The algorithm is now described in more detail.

2.1 Defining the Distribution of Interest

The distribution of interest (DI) specifies covariate locations at which the user wishes to make predictions, and it is one feature of the PBMSM that distinguishes it from standard procedures. The DI summarizes where the user wishes to make future predictions, and the new method uses this information to influence the model selection procedure. Different forms may be relevant for different studies, and the choice of a DI is flexible and situation specific.

Consider the chemical shipment example on page 253 of [17] (called Example 1). Here,

Y = the time (in minutes) to handle a shipment of drums

X_1 = the number of drums in the shipment

X_2 = the total weight of the shipment (in hundreds of pounds).

The collection of models considered is all subsets ($2^3 = 8$ models) of a full model, with two main effects and a two-factor interaction. If the warehouse manager is expecting larger-than-typical shipments in terms of the number of drums, then the goal is to use a Bayesian regression model that accurately predicts the time to handle these new shipments. The new shipments are assumed to have between 25 and 30 drums, but with unknown weight *a priori*. A first step in defining a DI is to examine the empirical relationship between covariates. The observed (X_1, X_2) pairs are presented in Figure 1, with "o" symbols. We see a positive linear relationship between X_1 and X_2 from Figure 1. Thus, a simple linear regression model describes plausible values of X_2 for a given value of X_1, with the distribution of X_2 conditional on X_1 assumed to be normal. This gives

$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi(3.96)}} \exp\left\{\frac{-1}{2(3.96)}(x_2 + 1.06 - 0.85x_1)^2\right\}. \qquad (2)$$

Using the assumed sizes of the new shipments, a natural choice for the marginal distribution of X_1 is the uniform distribution on the integers $\{25, 26, \cdots, 30\}$, denoted by $f(x_1)$. The joint probability density is the product of $f(x_2|x_1)$ and $f(x_1)$, which defines the DI. Figure 1 also depicts a sample from $f(x_2|x_1)f(x_1)$. In Figure 1, the "x" symbols mark the covariate locations sampled from the DI, and the line is the least squares regression line.

To characterize the empirical relationship between X_1 and X_2 to define the DI, the method of least squares was used. The use of least squares can be thought of as a heuristic to define the DI. In some situations, the user of the PBMSM may not need statistical methods to define the DI, as the DI may be naturally defined by the problem context. Such an example is presented in Section 4. Finally, for Example 1, the observed relationship between X_1 and X_2 is assumed to hold outside of the data range. One should always be

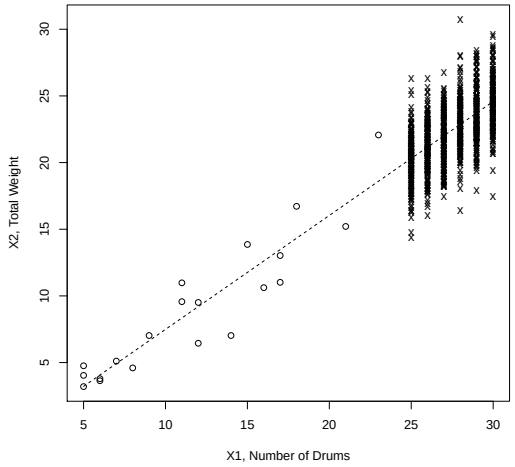


Figure 1: The "o" symbols represent the observed (X_1, X_2) pairs in Example 1 on page 253 of [17]. The x's represent a random sample of points from the DI in Example 1.

cautious with such an assumption, and whenever possible, base it on available underlying science. If that assumption is not true, predictions made at new covariate locations are unlikely to match the true process.

2.2 Sampling From the Distribution of Interest

The DI defines covariate locations at which predictions are likely to be sought. The goal is to select a model that predicts well over the entire DI. To assess a model's prediction ability over the entire DI, one must evaluate the prediction ability of each model at many locations. Randomly sampling from the DI provides representative coverage of the locations of interest. For Example 1 we first draw a random sample of $X_1$from $f(x_1)$. Then, for each sampled point x_1, we draw a random sample of X_2 (given $X_1 = x_1$) from $f(x_2|x_1)$. The sample size, N_{new}, should adequately cover the DI, but not be too large to be computationally infeasible.

Another natural DI might be a uniform distribution over a non-rectangular region. An example in two dimensions might be a non-rectangular parallelo-

gram. A rejection algorithm can be constructed to sample from such a DI, see page 253 of [5]. To sample uniformly from a non-rectangular parallelogram, first we sample from the smallest rectangle containing the parallelogram, and then keep only points inside the parallelogram.

2.3 Model Estimation and Evaluation

Model estimation and evaluation are discussed separately here because they present different challenges. For model estimation, the computational cost of estimating a large number of models can present challenges. For evaluation, finding a reasonable surrogate for the ideal result requires special consideration.

2.3.1 Model Estimation

The general form of the BMRM is given in (1). Model estimation in the Bayesian paradigm calculates $p^m(\boldsymbol{\beta}^m, \sigma_m^2|\boldsymbol{y})$ (the posterior distribution) from the observed y_i's using (1). For some simple forms of g^m, the posterior is available in closed form, and one such form of g^m is

$$g^m(\boldsymbol{\beta}^m, \sigma_m^2) = \frac{1}{\sigma_m^2};\ \sigma_m^2 \in (0, \infty);\ \boldsymbol{\beta}^m \in \mathbb{R}^{D_m}, \tag{3}$$

with D_m being the dimension of $\boldsymbol{\beta}^m$. In [11], the form of g^m in (3) is referred to as the standard non-informative prior distribution, and they state that $p^m(\boldsymbol{\beta}^m, \sigma_m^2|\boldsymbol{y})$ is proper under this prior when the model matrix is of full rank and the sample size is larger than D_m. Throughout this article, it is assumed that g^m is given by (3); however, a great strength of the Bayesian paradigm is its ability to leverage additional information, so if this is available, it should be included through the prior g^m. The closed form expression for p^m under the prior form in (3) is given by

$$p^m(\boldsymbol{\beta}^m, \sigma_m^2|\boldsymbol{y}) = p^m(\boldsymbol{\beta}^m|\sigma_m^2, \boldsymbol{y})p^m(\sigma_m^2|\boldsymbol{y}), \tag{4}$$

where

$$p^m(\boldsymbol{\beta}^m|\sigma_m^2,\boldsymbol{y}) = (2\pi\sigma_m^2)^{-D_m/2}|V_m|^{-1/2}\exp\left\{\frac{1}{2\sigma_m^2}(\boldsymbol{\beta}^m-\widehat{\boldsymbol{\beta}^m})'V_m^{-1}(\boldsymbol{\beta}^m-\widehat{\boldsymbol{\beta}^m})\right\}, \tag{5}$$

and

$$p^m(\sigma_m^2|\boldsymbol{y}) = \frac{(\nu_m/2)^{(\nu_m/2)}}{\Gamma(\nu_m/2)}s_m^{\nu_m}(\sigma_m^2)^{-(\nu_m/2+1)}\exp\left\{\frac{-\nu_m s_m^2}{2\sigma_m^2}\right\}. \tag{6}$$

Note that (5) is a multivariate normal distribution with mean vector $\widehat{\boldsymbol{\beta}^m}$, and variance-covariance matrix $\sigma^2 V_m$, with

$$\widehat{\boldsymbol{\beta}^m} = [(\boldsymbol{X}^m)'\boldsymbol{X}^m]^{-1}(\boldsymbol{X}^m)'\boldsymbol{y}, \tag{7}$$

and

$$V_m = [(\boldsymbol{X}^m)'\boldsymbol{X}^m]^{-1}, \tag{8}$$

where

$$\boldsymbol{X}^m = (\boldsymbol{x}_1^m, \boldsymbol{x}_2^m, \cdots, \boldsymbol{x}_n^m)' \tag{9}$$

We assume all of the models are parametrized so that $[(\boldsymbol{X}^m)'(\boldsymbol{X}^m)]$ is non-singular. For (6),

$$\nu_m = n - D_m, \tag{10}$$

and

$$s_m^2 = \frac{1}{\nu_m}(\boldsymbol{y}-\boldsymbol{X}^m\widehat{\boldsymbol{\beta}^m})'(\boldsymbol{y}-\boldsymbol{X}^m\widehat{\boldsymbol{\beta}^m}). \tag{11}$$

Although (4) has a convenient closed form, it is still complicated. Thus, quantities of interest are estimated using N_{samp} samples from (4). Sampling from (4) is straightforward because (5) and (6) are multivariate normal and inverse scaled χ^2 pdf's, respectively. Markov chain Monte Carlo (MCMC) methods are not required for the prior form in (3). In the model selection setting, non-informative flat priors are generally applicable because often little is known about the magnitude of the terms or which terms are likely to be active. However, the Bayesian approach allows us to incorporate prior knowledge if such knowledge exists, with conjugate priors for $\boldsymbol{\beta}^m$ and σ_m^2 of the form

$\boldsymbol{\beta}^m \sim N(\boldsymbol{q}, \sigma_m^2 \boldsymbol{R})$ and $\frac{ab}{\sigma_m^2} \sim \chi_a^2$.

In the preceding paragraph, methods for sampling from the posterior distribution of $\boldsymbol{\beta}^m$ are discussed, but these need to be converted to the posterior distribution for $\mu^m(\boldsymbol{x}_{new}^m) = E[y(\boldsymbol{x}_{new}^m)|\boldsymbol{\beta}^m] = (\boldsymbol{\beta}^m)'\boldsymbol{x}_{new}^m$. We use, $\boldsymbol{x}_{new}$ to generically refer to a covariate location sampled from the DI, and $\boldsymbol{x}_{new}^m$ to refer to the vector of predictor terms for model m and covariate location $\boldsymbol{x}_{new}$. The sample from the posterior distribution of $\boldsymbol{\beta}^m$ is transformed to give a sample from the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$, which is used to approximate properties of the posterior distribution. For each of the N_{new} covariate locations sampled from the DI, a posterior distribution is approximated for all N_{mod} models. To reduce the computational burden, a single sample of N_{samp} $\boldsymbol{\beta}^m$'s from (4) is used to explore all N_{new} posterior distributions. To ensure that this is reasonable in Example 1, the PBMSM was carried out with $N_{samp} = 5,000$ and $N_{samp} = 10,000$, and the results were compared.

2.3.2 Approximating $\mu(\boldsymbol{x}_{new})$

The ideal value, $\mu(\boldsymbol{x}_{new})$, is the true, but unknown, value of the quantity being predicted at covariate location $\boldsymbol{x}_{new}$. In the shipment example, $\mu(\boldsymbol{x}_{new})$ is the mean time to process a shipment of $x_{1,new}$ drums weighing $100 * x_{2,new}$ pounds. To judge the prediction abilities of the models under consideration, a reasonable surrogate for the ideal value, $\hat{\mu}(\boldsymbol{x}_{new})$, is calculated. This is important because some models may lead to posterior distributions of $\mu^m(\boldsymbol{x}_{new}^m)$ with small spreads but large biases, while others may lead to posterior distributions with large spreads and small biases. The surrogate is necessary to judge the bias of a model. Because *a priori* it is unknown which models under consideration lead to good point estimates of $\mu(\boldsymbol{x}_{new})$, BMA [21] is used because it combines information from all models weighted by the estimated qualities of the models. A point prediction, $\hat{\mu}^m(\boldsymbol{x}_{new}^m)$ (say the posterior mean), is taken from each model, and all the predictions are combined through weighted av-

eraging based on the posterior probabilities of the models. Specifically,

$$\hat{\mu}(\boldsymbol{x}_{new}) = \sum_{i=1}^{N_{mod}} w_i \hat{\mu}^i(\boldsymbol{x}^i_{new}), \tag{12}$$

where $\sum_{i=1}^{N_{mod}} w_i = 1$. In the case that model parameters are assigned proper priors, w_i will be the posterior probability of model i, $P(M = i|\boldsymbol{y})$. Here, model parameters are assigned improper priors, so $P(M = i|\boldsymbol{y})$ is not well defined, see pages 64 and 65 of [19].

As a substitute for a model's posterior probability when improper priors are employed, we use the approximation given by [20] on page 145:

$$P(M = m|\boldsymbol{y}) \approx \frac{\exp\{\frac{-1}{2}\text{BIC}_m\}\pi_m}{\sum_{i=1}^{N_{mod}} \exp\{\frac{-1}{2}\text{BIC}_i\}\pi_i}, \tag{13}$$

where BIC_m is the Bayesian information criterion (BIC) for model m [23] and π_m is the prior probability associated with model m. Note that when π_m is constant for all m, the π_m in the numerator and denominator will cancel. Because (13) is not a function of the prior distributions of the model parameters, it can be used in our situation. Equation (13) is derived by approximating the marginal pdf of $\boldsymbol{y}$ for model m, say $p(\boldsymbol{y}|M = m)$, as

$$p(\boldsymbol{y}|M = m) \approx \exp\{\frac{-1}{2}\text{BIC}_m\}. \tag{14}$$

This approximation uses the Laplace method for integrals and the asymptotic likelihood theory. See [20] for more details on the derivation.

If model parameters are assigned proper prior distributions, an MCMC algorithm can be used to sample realizations from the joint posterior distribution of $(\boldsymbol{\beta}^m, \sigma^2_m, m)$. Four general purpose algorithms are proposed by [4], [7], [13], and [15]. In [21], the methodology of [15] is used for linear models.

Again, the reader might feel that since we are comfortable using the model averaged point estimate as a surrogate for the true value over the DI, we should not continue with our process to select only one model. To answer this, we

remind the reader that our goal from the beginning has been to select a single model. We believe this is reasonable because practitioners of statistics, e.g. scientists and engineers, may prefer to work with a single model which lends itself more easily to interpretation than model averaging. We understand the advantages of model averaging, and we are not suggesting that our procedure is a replacement for it. We are simply leveraging the power of model averaging into our selection procedure.

2.3.3 The Discrepancy Measure

In Section 2.3.1, sampling from the posterior distribution of $\mu^m(\boldsymbol{x}^m_{new})$ at each $\boldsymbol{x}_{new}$ was discussed. In Section 2.3.2, approximating $\mu(\boldsymbol{x}_{new})$ at each $\boldsymbol{x}_{new}$ was considered. We now use these results to define a measure of prediction ability. For prediction in the Bayesian paradigm, the optimal scenario occurs when the posterior distribution is a point mass at the ideal value, $\mu(\boldsymbol{x}_{new})$, that is, when the posterior distribution provides unbiased predictions with no uncertainty. So to judge the prediction ability of model m at $\boldsymbol{x}_{new}$, the posterior distribution of $\mu^m(\boldsymbol{x}^m_{new})$ is compared to a point mass at $\hat{\mu}(\boldsymbol{x}_{new})$. If model m_1 is more similar to a point mass at $\hat{\mu}(\boldsymbol{x}_{new})$ than model m_2, model m_1 has better predictive ability than model m_2.

Let $F^m_{\boldsymbol{x}^m_{new}}$ denote the posterior cdf associated with $\mu^m(\boldsymbol{x}^m_{new})$ and $F_{\boldsymbol{x}_{new}}$ denote the cdf for a point mass at $\hat{\mu}(\boldsymbol{x}_{new})$. So $F_{\boldsymbol{x}_{new}}$ is a step function jumping from 0 to 1 at $\hat{\mu}(\boldsymbol{x}_{new})$. A natural way to compare these cdf's is by integrating their absolute difference, which is similar to the L_1 distance between two functions, see page 90 of [2]. Let

$$D_m(\boldsymbol{x}^m_{new}) = \int_{-\infty}^{\infty} |F^m_{\boldsymbol{x}^m_{new}}(u) - F_{\boldsymbol{x}_{new}}(u)| du. \tag{15}$$

Equation (15) generalizes to

$$D^k_m(\boldsymbol{x}^m_{new}) = \left\{ \int_{-\infty}^{\infty} |F^m_{\boldsymbol{x}^m_{new}}(u) - F_{\boldsymbol{x}_{new}}(u)|^k du \right\}^{\frac{1}{k}}, \tag{16}$$

which is similar to the L_k distance between two functions for $k \in [1, \infty)$. One should note that $D_m^k(\boldsymbol{x}_{new}^m)$ is almost identical to L_k in [2] on page 90 except that $\int |F_{\boldsymbol{x}_{new}^m}^m(u)|du = \infty$ and $\int |F_{\boldsymbol{x}_{new}}(u)|du = \infty$. The definition of L_k in [2] requires that $\int |F_{\boldsymbol{x}_{new}^m}^m(u)|du < \infty$ and $\int |F_{\boldsymbol{x}_{new}}(u)|du < \infty$, so that finite distance is guaranteed. However, $D_m^k(\boldsymbol{x}_{new}^m)$ is finite under the minimal conditions that the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ has finite expected value and $k \geqslant 1$. For a proof, see pages 80 and 81 of [19].

To help understand the metric, Figure 2 graphically depicts D_m under four different scenarios where "D" in the legend is the value of D_m, and the area of the gray shading lines graphically represents D_m. In all four scenarios, the step is at $\hat{\mu}(\boldsymbol{x}_{new}) = 0$. The upper left graphic depicts a scenario where the expected value of the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ matches $\hat{\mu}(\boldsymbol{x}_{new}) = 0$ well, and the spread of the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ is small. The bottom left graphic depicts a scenario where the expected value of the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ is shifted from 0, but the spread is still small. The upper right graphic depicts a scenario where the expected value of the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ matches $\hat{\mu}(\boldsymbol{x}_{new}) = 0$ well, but the spread is large. The bottom right graphic depicts a scenario where the expected value of the posterior distribution of $\mu^m(\boldsymbol{x}_{new}^m)$ is shifted from 0, and the spread is large. Note that an expected value close to $\hat{\mu}(\boldsymbol{x}_{new})$ with small spread leads to the smallest value of D_m.

Taking k to be small is recommended because as k increases, $D_m^k(\boldsymbol{x}_{new}^m)$ tends to either $F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})]$ or $1 - F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})]$. To see why, consider the following argument. Note that

$$\lim_{k\to\infty} \left[\int_{\mathbb{R}} |F_{\boldsymbol{x}_{new}^m}^m(u) - F_{\boldsymbol{x}_{new}}(u)|^k du \right]^{\frac{1}{k}} = \sup_{u\in\mathbb{R}} \{ |F_{\boldsymbol{x}_{new}^m}^m(u) - F_{\boldsymbol{x}_{new}}(u)| \} =$$
$$= \begin{cases} F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})] & \text{if } F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})] \geq 0.5 \\ 1 - F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})] & \text{if } F_{\boldsymbol{x}_{new}^m}^m[\hat{\mu}(\boldsymbol{x}_{new})] < 0.5 \end{cases}$$

because $F_{\boldsymbol{x}_{new}}$ is a step function at $\hat{\mu}(\boldsymbol{x}_{new})$. Thus, $D_m^k(\boldsymbol{x}_{new}^m)$ reduces to a single property of $F_{\boldsymbol{x}_{new}^m}^m$ as k increases. Because other important properties

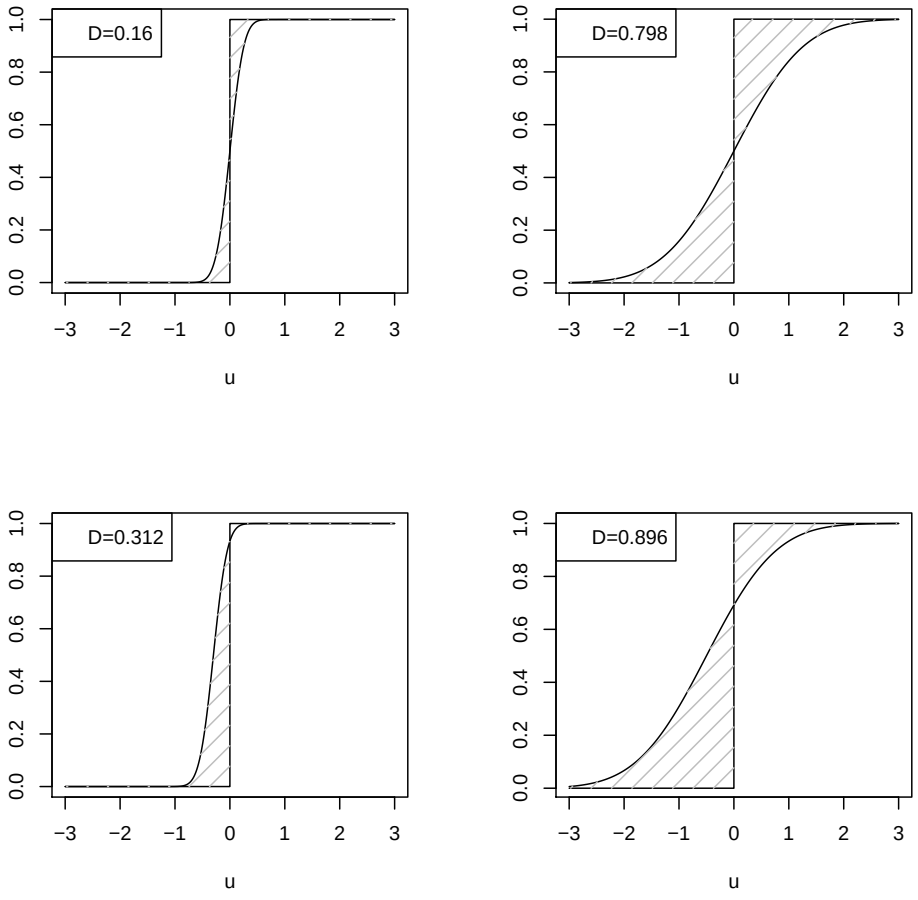


Figure 2: Graphical representations of D_m^1 under four circumstances. The area depicted by the gray diagonal shading lines represents D_m^1 where "D" in the legend is D_m^1.

of $F^m_{\boldsymbol{x}^m_{new}}$ are ignored, that reduction is undesirable. For instance, if $F^m_{\boldsymbol{x}^m_{new}}$ has large spread, the prediction from $F^m_{\boldsymbol{x}_{new}}$ has low precision. However, as k increases, the spread is progressively ignored. So we select $k = 1$, the lower boundary, in the remainder of this article. In Section 3, we see that with $k = 1$ the PBMSM outperforms other methods in terms of prediction ability. However, a clearly optimal setting for k in all scenarios is not available, but the sensitivity of the PBMSM to k in a given scenario can be assessed.

2.4 Comparing Models Graphically

Because $D_m(\boldsymbol{x}^m_{new})$ is estimated for $m = 1, 2, \cdots, N_{mod}$ and at each $\boldsymbol{x}_{new}$, the goal is to select the model with the most desirable distribution of discrepancy measures, D's. This is easy with a single location for the DI, as the model with the smallest D is selected. However, for most practical situations where the DI includes many (likely infinitely many) possible covariate locations, selecting the most desirable distribution of D's is not straightforward. Several different aspects of the distribution of D's could be used; one could focus on the average, median, maximum, or any percentile of the sampled D's. However, we recommend a graphical approach to simultaneously examine many characteristics.

First, using the boxplot of discrepancy measures for each model across all sampled locations from the DI, gross distinctions between models can be made. Figure 3(a) contains the set of boxplots for Example 1 where model 7 (X_1 and X_2) is clearly the best because small values for D are preferred.

In other scenarios, boxplots may not be sufficient to discern among models with similar performances. In these instances, we recommend fraction of covariate distribution (FCD) plots. FCD plots [19] are similar to fraction of design space (FDS) plots introduced by [25]. FCD plots are made by plotting the ordered discrepancy measures for each model on the vertical axis against $\frac{1}{N_{new}}, \frac{2}{N_{new}}, \cdots, \frac{N_{new}-1}{N_{new}}$, and 1 on the horizontal axis. More specifically, let $D_{m,(i)}$, $i = 1, 2, \cdots, N_{new}$, be the ith smallest discrepancy measure for model m. Then, the points $(\frac{i}{N_{new}}, D_{m,(i)})$ $i = 1, 2, \cdots, N_{new}$, are plotted. Figure 3(b)

contains the FCD curves for the four most competitive models, 3 (X_2 only), 5 (X_1 only), 7 (X_1 and X_2), and 8 (the full model). From Figure 3(b), it is again clear that model 7 is the best because an ideal curve is low and flat.

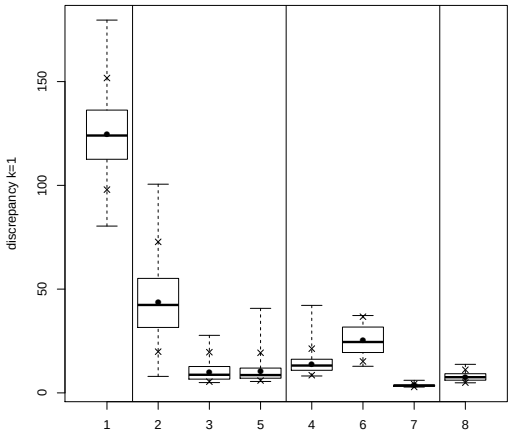
We now consider results for three other model selection methods. The DIC's and posterior probabilities for all eight models are presented in Table 1. Model 7 (X_1 and X_2) has the smallest DIC and is the highest posterior probability model (HPPM). The third method is referred to as the median probability model (MPM) in [3], and is also derived from the posterior probabilities. The first step in finding the MPM is calculating $P(\text{term } i \text{ is in the true model}|\boldsymbol{y}) = \sum_{m \in B_i} P(M = m|\boldsymbol{y})$, where B_i is the set of models that contain term i. Each term with a posterior probability of inclusion over 0.5 is considered to be important. For Example 1, the values are 1, 1, and 0.24 for X_1, X_2, and X_1X_2, respectively, with model 7 again highlighted as the best.

Model	Terms	Posterior Probability	DIC
1	None	0	213.18
2	X_1X_2	0	162.77
3	X_2	0	151.82
4	X_2, X_1X_2	0	153.37
5	X_1	0	158.32
6	X_1, X_1X_2	0	148.58
7	X_1, X_2	0.76	130.65
8	X_1, X_2, X_1X_2	0.24	132.19

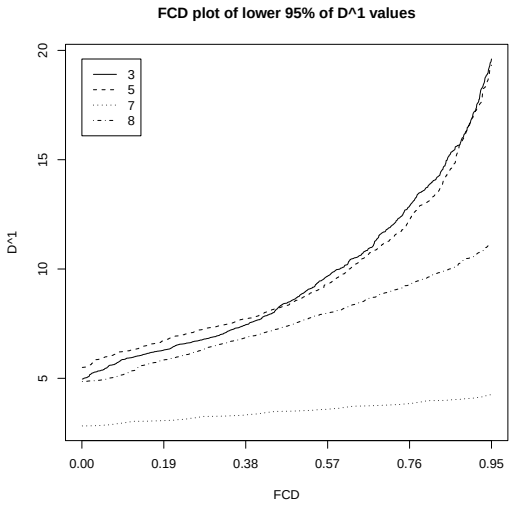
Table 1: Posterior probabilities and DIC's for all eight models in Example 1.

3 Simulation Study

In this section, a simulation study compares the new model selection method to three other methods. The three methods select the HPPM, the MPM, and the model with the lowest DIC. The metric to evaluate the methods is based on prediction ability. More specifically, the distance of a selected model's prediction from the true value is evaluated because the true data generating model is known.



(a)



(b)

Figure 3: (a) Boxplots of discrepancy measures in Example 1. (b) FCD plots in Example 1 for models 3 (X_2 only), 5 (X_1 only), 7 (X_1 and X_2), and 8 (the full model).

In the simulation study, 45 distinct scenarios are considered. The factors distinguishing the scenarios are the correlation level between the predictors, the DI, and the true model. Each scenario involves two covariates, X_1 and X_2, and eight models corresponding to all subsets of a full model with both main effects and the two-factor interaction. For each scenario, 30 data points are observed, and $N_{new} = 1,000$ covariate locations are sampled from the DI. Finally, $N_{sim} = 2,000$ data sets are simulated and analyzed for each scenario.

The different scenarios form a full factorial with three factors: correlation level (3 levels), DI (3 levels), and true model (5 levels). The three correlation levels are 0, 0.8, and 0.95. The three DI's characterize an extrapolated region (labeled DI 1), the entire observed data region (DI 2), and only a portion of the observed data region (DI 3). These nine combinations of correlation level and DI are illustrated in Figure 4 where the o's represent observed covariate points, and the x's represent sampled locations from the DI's. For all combinations, five true models are examined. Those models are $\mu(\boldsymbol{x}) = x_1$, $\mu(\boldsymbol{x}) = 2x_1$, $\mu(\boldsymbol{x}) = x_1 + x_2$, $\mu(\boldsymbol{x}) = 2x_1 + 2x_2$, and $\mu(\boldsymbol{x}) = 2x_1 + x_2 + x_1x_2$. The "*new*" subscript is omitted because the true model applies to both the new locations sampled from the DI and the observed covariate points. All true models use $\sigma^2 = 1$.

For each scenario, the simulation uses these steps:

1. Generate a data set from the true model based on a fixed set of $\boldsymbol{x}_i$'s for that scenario.

2. Perform model selection with each of the four methods where models are assumed to be equally likely *a priori*, and the model parameters are assigned the standard improper priors.

3. Quantify the prediction error, $E_m = \frac{1}{N_{new}} \sum_{i=1}^{N_{new}} [\widehat{\mu}^m(\boldsymbol{x}^m_{new,i}) - \mu_{true}(\boldsymbol{x}_{new,i})]^2$, of each selected model where m is one of m_{PBMSM}, m_{HPPM}, m_{MPM}, or m_{DIC} selected by its respective model selection algorithm, and $\widehat{\mu}^m(\boldsymbol{x}^m_{new,i})$ is the mean of the posterior distribution of $\mu^m(\boldsymbol{x}^m_{new,i})$ for $\boldsymbol{x}_{new,i}$, the ith covariate location sampled from the DI.

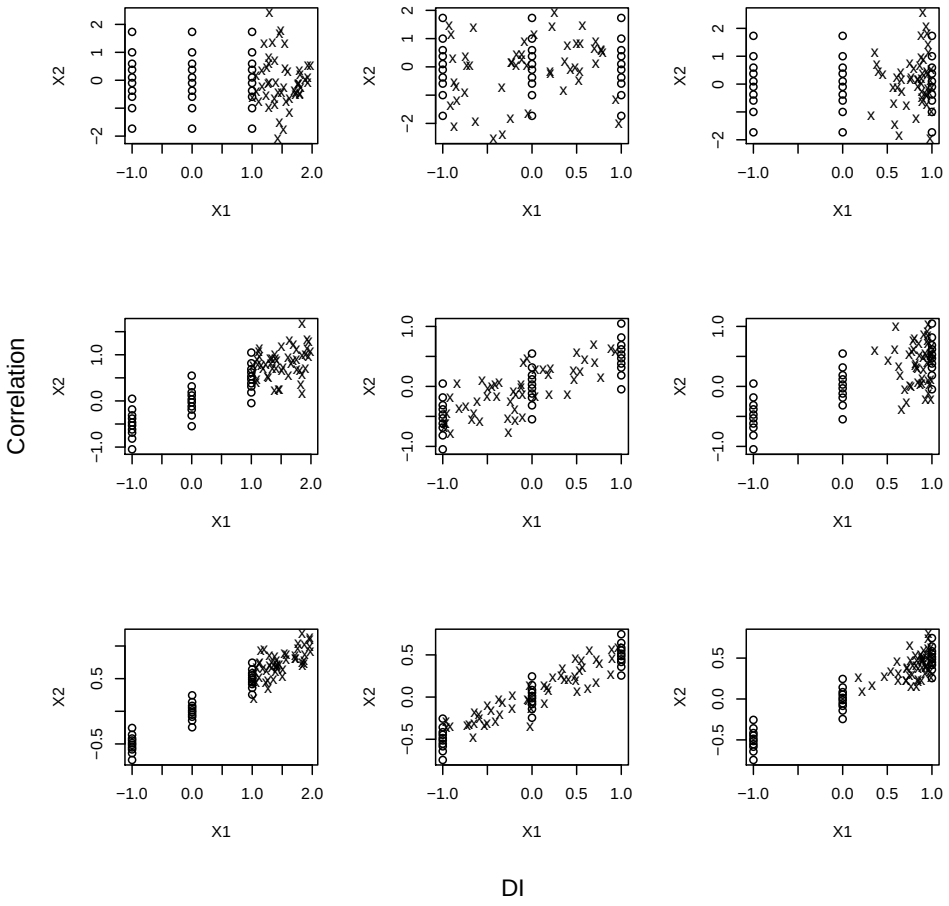


Figure 4: The nine combinations of correlation level and DI. The o's represent the observed covariate points and the x's represent covariate locations sampled from the DI.

The prediction error, E_m, is the average squared distance (over the points sampled from the DI) from a selected model's prediction to the true value. Because the goal of the simulation study is to assess the prediction ability of the four selection algorithms, E_m is a natural metric. If a selection algorithm leads to predictions close to the true value, E_m will be small. For each of the 45 scenarios, each selection methodology has $N_{sim} = 2,000$ simulation values, E_m's, which we summarize with the mean, median, and 90th and 95th percentiles.

Because the new model selection procedure is repeated many times across the many data sets and scenarios, graphical comparisons are not practical. (We still recommend graphical comparisons for an individual analysis because many considerations from the whole distribution of discrepancies can be factored into the selection decision.) An automated numerical summary of $\{D_m(\boldsymbol{x}^m_{new,i})|i = 1, 2, \cdots, N_{new}\}$ is used in the simulation study, where $q_\alpha(D_m)$ is its α percentile. Then, the summary of $\{D_m(\boldsymbol{x}^m_{new,i})|i = 1, 2, \cdots, N_{new}\}$ is $\zeta_m = \frac{1}{11}\sum_{j=1}^{11} q_{\alpha_j}(D_m)$ where $\alpha_1 = 0.05$, $\alpha_2 = 0.1$, $\alpha_3 = 0.2$, $\cdots$, $\alpha_{10} = 0.9$, and $\alpha_{11} = 0.95$. The model with the smallest value of ζ_m is used as the best for the new method. This can be thought of as comparing 11 distinct points that comprise FCD curves. As we expect, if the FCD curve for model m_1 is always below the FCD curve for model m_2, $\zeta_{m_1} < \zeta_{m_2}$.

Table 2 summarizes the values of E_m from the simulation study. Consider 0.5605 in the third row of the first column, which corresponds to the PBMSM. For each of the five models within the scenarios with DI 1 and correlation 0, the 90th percentile was calculated from the 2,000 data sets. The value 0.5605 is the average of the five 90th percentiles across the different true models.

Because smaller values of E_m are preferable, the row minimums of Table 2 are highlighted. It is clear that the PBMSM often out-performs the other selection methodologies; however, when the DI is 2 or 3 and the correlation is 0.8, selecting the model with the lowest DIC out-performs the PBMSM.

Table 2 summarizes the results of the entire simulation study, which include the scenarios that use the full model as the true model. In practice, the full model is often larger than necessary so some model reduction is anticipated.

	PBMSM	HPPM	MPM	DIC
Distribution=1; Correlation=0				
mean	**0.2322**	0.2368	0.2369	0.2578
median	**0.1256**	0.1315	0.1317	0.1547
90th	**0.5605**	0.5826	0.5831	0.6357
95th	**0.7990**	0.8184	0.8164	0.8595
Distribution=1; Correlation=0.80				
mean	**0.4306**	0.4435	0.4407	0.4508
median	0.2714	0.2630	0.2601	**0.2261**
90th	**0.9522**	1.0159	1.0118	1.1652
95th	**1.3840**	1.5152	1.5120	1.6737
Distribution=1; Correlation=0.95				
mean	**0.4906**	0.5159	0.5099	0.5629
median	0.2843	0.2858	0.2741	**0.2610**
90th	**0.9943**	1.1082	1.0992	1.4979
95th	**1.6208**	1.9402	1.9462	2.2610
Distribution=2; Correlation=0				
mean	**0.0883**	0.0906	0.0906	0.0971
median	**0.0637**	0.0658	0.0659	0.0754
90th	**0.1907**	0.1978	0.1976	0.2058
95th	**0.2533**	0.2571	0.2569	0.2641
Distribution=2; Correlation=0.80				
mean	0.1227	0.1235	0.1210	**0.1169**
median	0.0928	0.0941	0.0928	**0.0914**
90th	0.2613	0.2555	0.2508	**0.2424**
95th	0.3440	0.3443	0.3420	**0.3176**
Distribution=2; Correlation=0.95				
mean	**0.1089**	0.1159	0.1108	0.1131
median	**0.0800**	0.0823	**0.0800**	0.0835
90th	**0.2389**	0.2605	0.2459	0.2523
95th	**0.3006**	0.3195	0.3097	0.3138
Distribution=3; Correlation=0				
mean	**0.1123**	0.1139	0.1136	0.1219
median	**0.0657**	0.0678	0.0678	0.0805
90th	**0.2682**	0.2715	0.2707	0.2893
95th	**0.3696**	0.3716	0.3702	0.3873
Distribution=3; Correlation=0.80				
mean	0.1385	0.1430	0.1402	**0.1318**
median	**0.0887**	0.0917	0.0896	0.0890
90th	0.3301	0.3395	0.3338	**0.3107**
95th	0.4068	0.4130	0.4105	**0.4010**
Distribution=3; Correlation=0.95				
mean	**0.1197**	0.1271	0.1224	0.1216
median	**0.0815**	0.0856	0.0823	0.0838
90th	**0.2711**	0.2887	0.2786	0.2793
95th	**0.3542**	0.3727	0.3654	0.3627

Table 2: Summary of results from the simulation study.

Table 3 summarizes the results of the simulation study, but it excludes the scenarios for which the full model is the true model. Table 3 shows that the PBMSM is either best or very competitive in all of the considered scenarios.

4 An Example Involving a Large Number of Potential Models

In Example 1 and the simulation study, a small number of competing models was considered. The second example (Example 2) below considers a far larger collection of competing models, which may be more typical of many practical situations.

4.1 Introduction

The data set on page 596 of [8] involves a designed experiment using a central composite design (CCD) with axial values of $\alpha = 2$, see page 49 of [16]. The design considers four factors:

X_1 = percentage of H_2O_2 by weight of paper,
X_2 = percentage of NaOH by weight of paper,
X_3 = percentage of silicate by weight of paper,
X_4 = process temperature.

The response, Y, is the brightness of finished paper, and the full model is a full quadratic model with 4 linear effects, 4 pure quadratic terms, and 6 two-way interactions. The competing models are taken to be all subsets of the full model. We assume that the goal of the experiment is to develop a model that predicts well over a small hypercube near the boundary of the design region.

4.2 Selecting and Sampling From the Distribution of Interest

To meet the goal of the study, a DI is constructed as a uniform distribution on the four dimensional hypercube, $[1, 1.5]^4$. Sampling a covariate lo-

	PBMSM	HPPM	MPM	DIC
Distribution=1; Correlation=0				
mean	**0.2103**	0.2196	0.2197	0.2482
median	**0.1073**	0.1147	0.1149	0.1436
90th	**0.5270**	0.5558	0.5565	0.6229
95th	**0.7632**	0.7958	0.7933	0.8550
Distribution=1; Correlation=0.80				
mean	**0.2944**	0.3205	0.3174	0.3764
median	**0.1392**	0.1456	0.1419	0.1668
90th	**0.6634**	0.7484	0.7442	0.9952
95th	**1.1000**	1.2690	1.2674	1.5217
Distribution=1; Correlation=0.95				
mean	**0.3355**	0.3735	0.3704	0.4720
median	**0.1267**	0.1371	0.1285	0.1575
90th	**0.6749**	0.8109	0.8062	1.3437
95th	**1.3406**	1.7373	1.7483	2.1641
Distribution=2; Correlation=0				
mean	**0.0821**	0.0851	0.0851	0.0936
median	**0.0569**	0.0595	0.0597	0.0715
90th	**0.1839**	0.1931	0.1929	0.2036
95th	**0.2445**	0.2506	0.2504	0.2612
Distribution=2; Correlation=0.80				
mean	**0.1053**	0.1092	0.1061	0.1082
median	**0.0721**	0.0757	0.0741	0.0805
90th	0.2481	0.2429	0.2372	**0.2343**
95th	0.3351	0.3374	0.3345	**0.3115**
Distribution=2; Correlation=0.95				
mean	**0.1003**	0.1070	0.1028	0.1077
median	**0.0742**	0.0771	0.0747	0.0805
90th	**0.2194**	0.2394	0.2274	0.2390
95th	**0.2796**	0.2971	0.2899	0.3005
Distribution=3; Correlation=0				
mean	**0.1050**	0.1077	0.1074	0.1181
median	**0.0587**	0.0614	0.0614	0.0773
90th	**0.2550**	0.2602	0.2593	0.2827
95th	**0.3549**	0.3602	0.3584	0.3804
Distribution=3; Correlation=0.80				
mean	**0.1235**	0.1274	0.1244	**0.1235**
median	**0.0759**	0.0793	0.0775	0.0825
90th	0.3108	0.3201	0.3137	**0.2939**
95th	**0.3885**	0.3927	0.3897	0.3896
Distribution=3; Correlation=0.95				
mean	**0.1159**	0.1218	0.1182	0.1186
median	**0.0777**	0.0809	0.0783	0.0814
90th	**0.2683**	0.2819	0.2741	0.2745
95th	**0.3470**	0.3622	0.3589	0.3558

Table 3: Summary of results from the simulation study, but scenarios for which the full model is the true model are excluded.

cation, $\boldsymbol{x}_{new} = (x_{new,1}, x_{new,2}, x_{new,3}, x_{new,4})'$, is done by sampling $x_{new,j}$ for $j = 1, 2, 3, 4$ independently from a uniform distribution on $[1, 1.5]$.

The relationship between the covariates is one of two major differences between Examples 1 and 2. Selecting and sampling from the DI in Example 2 is straightforward because the covariates arise from a designed experiment and are uncorrelated. In general, selecting the DI is an important step in the PBMSM, and the choice should be based on the goal of the analysis, the observed relationship between the covariates, and advice from subject matter experts.

4.3 Calculation of D_m

If we consider all subsets of the full model, Example 2 has $2^{14} = 16,384$ models because the full quadratic model in four factors has 14 terms plus the intercept (which we assume is always included). It is preferable to reduce the number of competing models, if possible, before carrying out the PBMSM. One possibility for this is to consider only models adhering to strong or weak heredity principle [6]. These principles place conditions on the inclusion of interaction and quadratic terms, and this helps the scientific interpretability of the selected model. Both the strong and weak heredity principles can lead to large reductions in the number of competing models.

Another way to reduce the number of competing models is to realize that many of them will predict very poorly, and they can be quickly disregarded. To accomplish the reduction, only models with high posterior probabilities relative to the highest posterior probability are considered. The models in this reduced set belong to Occam's window (OW), and by this definition, Occam's window is said to be symmetric [20].

To find this reduced set of models, posterior probabilities are first approximated for all models using (13). Let $\mathcal{P}$ represent the collection of all approximated posterior probabilities, and let $M_{\mathcal{P}} = \max(\mathcal{P})$. Then, the reduced set is $OW = \{m | \frac{M_{\mathcal{P}}}{P(M=m|\boldsymbol{y})} < \omega\}$. The interpretation of ω is intuitive: If $\omega = 50$, the models not included in OW have posterior probabilities that

are less than $\frac{1}{50} = 2\%$ of $M_{\mathcal{P}}$. Taking $\omega = 50$ reduces the number of models from $16,384$ to 310, and taking $\omega = 10$ reduces the number of models to 53. Figure 5 displays the trade-offs between using $\omega = 10$ and $\omega = 50$ for Example 2 where increasing ω from 10 to 50 leads to 257 extra models in OW; however, the posterior probability of each of those 257 models is less than 0.005. In this example, the Occam's window approach (not the heredity principles) is used to reduce the number of competing models; however, it would be possible to use both approaches.

After reducing the set of competing models, we renormalize the posterior probabilities to make the probabilities of the remaining models sum to 1. Then, the calculation of D_m for all models in the reduced set is carried out as before. In Example 2, the values of ω, N_{samp}, and N_{new} are chosen as 50, $5,000$, and $1,000$, respectively.

4.4 Comparing Models

Graphical comparisons of 310 models (based on $\omega = 50$) still are not an easy task. Thus, the set of 310 models, for which discrepancy measures are calculated, is further reduced before graphical comparisons by examining a table of summary statistics for the best models.

Table 4 lists the 95th percentile and mean discrepancy measures for the five best models for each number of terms according to the 95th percentile. In Table 4 local rank refers to a model's rank within a specified number of terms, and global rank refers to a model's rank among all 310 models in OW. More than one summary statistic is used because we prefer to judge a model on its distribution of discrepancy measures, not on a single summary statistic. One might use other summary statistics besides the 95th percentile and the mean. Table 4 also lists the terms in each model. Since referring to a model by its terms is cumbersome here, the models will be referred to by their numeric names (given in parentheses in the second column of Table 4) in the remainder of this section.

From Table 4, the five overall best models with respect to the 95th per-

# terms	Model	95th Percentile value	95th Percentile rank	Mean value	Mean rank
2	X_1, X_2 (1)	0.47583	1(247)	0.40305	1(257)
3	X_1, X_2, X_1X_4 (5)	0.24162	1(1)	0.21226	1(2)
3	X_1, X_2, X_2X_3 (3)	0.24232	2(2)	0.21449	2(3)
3	X_1, X_2, X_3 (135)	0.31748	3(50)	0.27404	3(68)
3	X_1, X_2, X_4 (84)	0.34976	4(128)	0.2916	4(127)
3	X_1, X_2, X_1^2 (32)	0.62267	5(292)	0.51453	5(292)
4	X_1, X_2, X_4, X_2X_3 (86)	0.2579	1(5)	0.21868	2(6)
4	X_1, X_2, X_3, X_2X_3 (137)	0.26632	2(8)	0.21695	1(4)
4	X_1, X_2, X_3, X_1X_4 (139)	0.2761	3(11)	0.22111	4(9)
4	X_1, X_2, X_4^2, X_1X_4 (21)	0.28052	4(13)	0.23676	5(14)
4	X_1, X_2, X_4, X_1X_4 (88)	0.28344	5(15)	0.22088	3(8)
5	X_1, X_2, X_3, X_1^2, X_1X_4 (175)	0.24902	1(3)	0.21151	1(1)
5	X_1, X_2, X_3, X_1^2, X_2X_3 (172)	0.25468	2(4)	0.21845	2(5)
5	X_1, X_2, X_4, X_1^2, X_1X_4 (104)	0.25906	3(6)	0.21945	3(7)
5	X_1, X_2, X_4, X_1^2, X_2X_3 (102)	0.26501	4(7)	0.22574	4(11)
5	X_1, X_2, X_1X_4, X_2X_3, X_2X_4 (10)	0.29921	5(24)	0.2542	6(32)
6	X_1, X_2, X_3, X_1^2, X_4^2, X_1X_4 (196)	0.27087	1(9)	0.22561	1(10)
6	X_1, X_2, X_1^2, X_4^2, X_1X_4, X_2X_3 (61)	0.27783	2(12)	0.23328	2(13)
6	X_1, X_2, X_3, X_4, X_1^2, X_2X_3 (257)	0.28802	3(16)	0.23913	4(17)
6	X_1, X_2, X_3, X_1^2, X_4^2, X_2X_3 (194)	0.29336	4(20)	0.23732	3(15)
6	X_1, X_2, X_3, X_4, X_1X_4, X_2X_4 (239)	0.29484	5(22)	0.25006	6(26)
7	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_4 (277)	0.27472	1(10)	0.23094	1(12)
7	X_1, X_2, X_3, X_1^2, X_1X_4, X_2X_3, X_2X_4 (180)	0.2808	2(14)	0.23863	2(16)
7	X_1, X_2, X_4, X_1^2, X_1X_4, X_2X_3, X_2X_4 (109)	0.29227	3(19)	0.24714	3(21)
7	X_1, X_2, X_1^2, X_3^2, X_1X_4, X_2X_3, X_2X_4 (70)	0.31272	4(42)	0.26579	6(52)
7	X_1, X_2, X_3, X_4^2, X_1X_4, X_2X_3, X_2X_4 (156)	0.31409	5(45)	0.26746	7(55)
8	X_1, X_2, X_3, X_1^2, X_3^2, X_1X_4, X_2X_3, X_2X_4 (213)	0.29081	1(18)	0.24676	1(19)
8	X_1, X_2, X_3, X_1^2, X_2^2, X_1X_4, X_2X_3, X_2X_4 (225)	0.29406	2(21)	0.24713	2(20)
8	X_1, X_2, X_4, X_1^2, X_3^2, X_1X_4, X_2X_3, X_2X_4 (126)	0.29764	3(23)	0.25085	5(28)
8	X_1, X_2, X_3, X_1^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (200)	0.30168	4(27)	0.24785	3(22)
8	X_1, X_2, X_4, X_1^2, X_2^2, X_1X_4, X_2X_3, X_2X_4 (131)	0.30321	5(31)	0.25371	6(31)
9	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (281)	0.28885	1(17)	0.24981	2(25)
9	X_1, X_2, X_3, X_1^2, X_3^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (218)	0.30197	2(28)	0.25132	3(29)
9	X_1, X_2, X_3, X_1^2, X_2^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (230)	0.3027	3(30)	0.24934	1(24)
9	X_1, X_2, X_3, X_4, X_1^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (269)	0.30831	4(36)	0.26547	6(48)
9	X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1X_4, X_2X_3, X_2X_4 (234)	0.31528	5(47)	0.26132	4(41)
10	X_1, X_2, X_3, X_4, X_1^2, X_3^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (295)	0.30853	1(37)	0.2667	4(54)
10	X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (305)	0.30991	2(41)	0.26474	3(47)
10	X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (235)	0.31405	3(44)	0.26105	1(40)
10	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (284)	0.31713	4(49)	0.26432	2(45)
10	X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (303)	0.3246	5(68)	0.27607	5(73)
11	X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_4^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (307)	0.32457	1(67)	0.27247	1(65)
11	X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (310)	0.32764	2(72)	0.27803	3(83)
11	X_1, X_2, X_3, X_4, X_1^2, X_3^2, X_4^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (297)	0.32852	3(74)	0.27645	2(75)
11	X_1, X_2, X_3, X_4, X_1^2, X_1^2, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (288)	0.34257	4(111)	0.28997	4(121)
11	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4 (285)	0.34401	5(118)	0.2924	5(130)

Table 4: The 95th percentile and mean discrepancy measures for the best models in Example 2. The rank within parentheses is a model's rank relative to all of the other models, and the rank outside of parentheses is a model's rank relative to all other models with the same number of terms.

centile or mean discrepancy are models 3, 5, 86, 137, 172, 175. From these six models, a winner is selected using graphical comparison. Figure 6 contains FCD curves of the discrepancy measures for these six models. It provides evidence that model 175 is preferred over the others because it has the smallest discrepancy measure for more than 50% of the DI, and for locations in the DI where models 3 and 5 have smaller discrepancy measures, model 175 is not much worse. Here, the PBMSM identifies two or three models that are best over different parts of the DI. We view this as a strength of the PBMSM because at this point, subject matter specialists can incorporate non-statistical criteria, such as the cost of future data collection, into the final selection process.

In comparison to the PBMSM, Table 5 displays the four models with the highest posterior probabilities and the four models with the lowest DIC values. There is considerable overlap because models 264, 180, and 281 appear in both lists as the top three models. The marginal posterior probabilities for each of the 14 terms are given in Table 6. Based on this, the MPM is model 180. Note that there is not much overlap between the three current methods and the PBMSM. This is due to the DI's dissimilarity to the design region. In contrast, if the DI were taken to be the hyper-cube $[-2, 2]^4$ (the hyper-cube containing the design region), there would be considerable overlap between the current methods and the PBMSM. Specifically, the best model from the PBMSM would be the same as the MPM, and three of the best models from the PBMSM would have a top four posterior probabilities and DIC's. With the DI being $[1, 1.5]^4$, the best model from the PBMSM is not the MPM, and none of the best models from the PBMSM have a top four posterior probability or DIC. Also note that the models selected as the best by the PBMSM (with the DI being $[1, 1.5]^4$) tend to have less terms than the models selected as the best by the current methods. This is expected because the DI is near the outer boundary of the design region where there is a greater penalty for predicting based on including additional terms as the spread of the posterior distribution of $\mu^m(\boldsymbol{x}^m_{new})$ is larger than that near the center of the design region.

We realize that the number of models considered moderate by today's

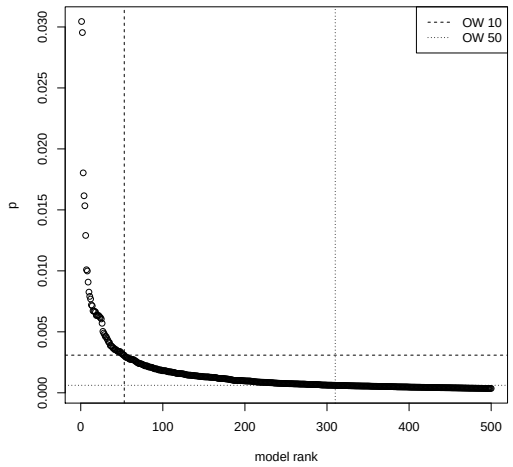


Figure 5: The trade-offs between using $\omega = 10$ and $\omega = 50$ for defining Occam's window in Example 2. The symbol "p" on the vertical axis stands for posterior probability. The horizontal axis is the model's rank according to posterior probability.

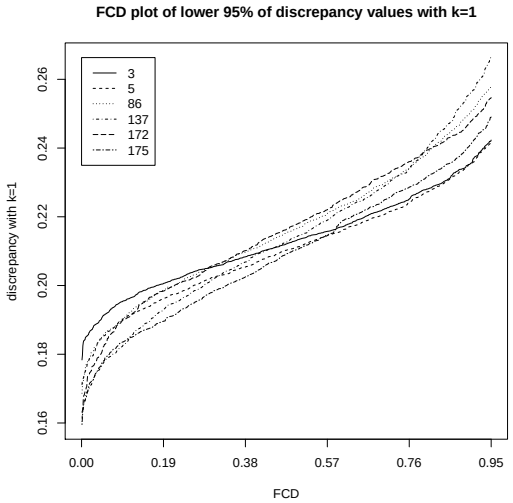


Figure 6: FCD plots of discrepancy measures for the best models in Example 2.

# Terms	Model	Posterior Probability
8	X_1, X_2, X_3, X_4, X_1^2, X_1X_4, X_2X_3, X_2X_4 (264)	0.0437
7	X_1, X_2, X_3, X_1^2, X_1X_4, X_2X_3, X_2X_4 (180)	0.0424
9	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (281)	0.0259
6	X_1, X_2, X_1^2, X_1X_4, X_2X_3, X_2X_4 (43)	0.0232
		DIC
8	X_1, X_2, X_3, X_4, X_1^2, X_1X_4, X_2X_3, X_2X_4 (264)	30.3746
9	X_1, X_2, X_3, X_4, X_1^2, X_4^2, X_1X_4, X_2X_3, X_2X_4 (281)	30.4310
7	X_1, X_2, X_3, X_1^2, X_1X_4, X_2X_3, X_2X_4 (180)	31.3797
9	X_1, X_2, X_3, X_4, X_1^2, X_1X_3, X_1X_4, X_2X_3, X_2X_4 (269)	31.4917

Table 5: Posterior probabilities and DIC's for the top models in Example 2.

Term	Posterior Probability
X_1	1.0000
X_2	1.0000
X_3	0.6392
X_4	0.4269
X_1^2	0.7587
X_2^2	0.1100
X_3^2	0.1100
X_4^2	0.2583
X_1X_2	0.1018
X_1X_3	0.1653
X_1X_4	0.9086
X_2X_3	0.8561
X_2X_4	0.7939
X_3X_4	0.1181

Table 6: Marginal posterior probabilities of each term in Example 2.

standards; however, this example illustrates nicely how to apply the algorithm when there are too many models under consideration for graphical comparisons alone.

5 Concluding Remarks

In this article, we developed a model selection algorithm for Bayesian linear models. The procedure, PBMSM, focuses on good prediction in a user-specified portion of the covariate space defined by the DI. The PBMSM uses a sequence of four steps. The first step is to define the DI. The second step samples from the DI. The third step calculates the discrepancy measure at each covariate location sampled from the DI for each model under comparison. The final step compares models graphically and numerically to identify a model or models which perform best for the DI.

Two examples are considered in this article. Example 1, which identifies a model to predict handling time for a shipment of drums, is used to illustrate the four steps of the procedure. It provides an instance when the covariates are naturally correlated and the DI naturally falls outside of the observed data range. Simple linear regression is employed to extend the observed relationship of the two covariates. The HPPM, the MPM, the model with the smallest DIC, and the model selected by the PBMSM all highlight the same best model.

Example 2 is from a designed experiment, and provides a new challenge of a large number ($2^{14} = 16,384$) of competing models. It serves to illustrate how these models can be reduced to a much smaller set of promising models before applying the PBMSM. Posterior model probabilities are calculated, and models with small posterior probabilities are excluded from further investigation. In Example 2, the PBMSM highlights models that tend to be smaller than the models highlighted by the three other standard methods. This difference is attributable to the selected DI because when the DI matches the design region more closely, the PBMSM produces results that are more in line with the three standard methods.

A simulation study is also presented. It considers a range of true models,

correlation levels between covariates, and DI's when two covariates are present. The simulation study shows that the PBMSM performs well in most of the considered scenarios. The largest improvement is seen when the DI is outside of the observed data range and correlation between the covariates is high. This is expected because in that situation, the variability of predictions is inflated. Thus, a smaller model with less variable predictions may be preferred.

The model selection procedure presented in this article focuses on good prediction in a user-specified portion of the covariate space. Many model selection procedures are available, with each developed for particular objectives. Users of model selection algorithms should consider those objectives when selecting an approach. We have provided evidence that when good prediction over a specific portion of the covariate space is the goal, the PBMSM is a good choice.

References

[1] H. Akaike, A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control 19 (1974), 716–723.

[2] K. B. Athreya and S. N. Lahiri, Measure Theory and Probability Theory, Springer, 2006.

[3] M. M. Barbieri and J. O. Berger, Optimal Predictive Model Selection, The Annals of Statistics 32 (2004), 870–897.

[4] B. P. Carlin and S. Chib, Bayesian Model Choice via Markov Chain Monte Carlo Methods, J. Royal Stat. Soc. Ser. B 57 (1995), 473–484.

[5] G. Casella and R. L. Berger, Statistical Inference, 2nd ed., Duxbury, 2002.

[6] H. Chipman, Bayesian Variable Selection with Related Predictors, Can. J. Stat. 24 (1996), 17–36.

[7] P. Dellaportas, J. J. Forster and I. Ntzoufras, On Bayesian Model and Variable Selection Using MCMC, Department of Statistics, Athens University of Economics and Business, Tech. Rep., 1998.

[8] J. L. Devore, Probability and Statistics for Engineering and the Sciences, 7th ed., Brooks/Cole, 2009.

[9] S. Geisser and W. F. Eddy, A Predictive Approach to Model Selection, J. Amer. Stat. Assoc. 74 (1979), 153–160.

[10] A. E. Gelfand and S. K. Ghosh, Model Choice: A Minimum Posterior Predictive Loss Approach, Biometrika 85 (1998), 1–11.

[11] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, Bayesian Data Analysis, Chapman & Hall/CRC, 2004.

[12] E. I. George and R. E. McCulloch, Variable Selection via Gibbs Sampling, J. Amer. Stat. Assoc. 88 (1993), 881–889.

[13] P. J. Green, Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, Biometrika 82 (1995), 711–732.

[14] P. W. Laud and J. G. Ibrahim, Predictive Model Selection, J. Royal Stat. Soc. Ser. B 57 (1995), 247–262.

[15] D. Madigan and J. York, Bayesian Graphical Models for Discrete Data, Inter. Stat. Rev. 63 (1995), 215–232.

[16] R. H. Myers, D. C. Montgomery and C. M. Anderson-Cook, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, John Wiley & Sons, 2009.

[17] J. Neter, M. H. Kutner, C. J. Nachtsheim and W. Wasserman, Applied Linear Statistical Models, 4th ed., WCB/McGraw-Hill, 1996.

[18] A. Ozol-Godfrey, C. M. Anderson-Cook and D. C. Montgomery, Fraction of Design Space Plots for Examining Model Robustness, J. Qual. Tech., 37 (2005), 223–235.

[19] A. L. Pintar, Model Selection for Good Estimation or Prediction Over a User-Specified Covariate Distribution, Ph.D. Dissertation, Iowa State University, Ames, Iowa, 2010.

[20] A. E. Raftery, Bayesian Model Selection in Social Research, Sociological Methodology 25 (1995), 111–163.

[21] A. E. Raftery, D. Madigan and J. A. Hoeting, Bayesian Model Averaging for Linear Regression Models, J. Amer. Stat. Assoc. 92 (1997), 170–191.

[22] A. San Martini and F. Spezzaferri, A Predictive Model Selection Criterion, J. Royal Stat. Soc. Ser. B 46 (1984), 296–303.

[23] G. Schwarz, Estimating the Dimension of a Model, The Annals of Statistics 6 (1978), 461–464.

[24] D. J. Spiegelhalter, N. G. Best, B. P. Carlin and A. Van Der Linde, Bayesian Measures of Model Complexity and Fit, J. Royal Stat. Soc. Ser. B 64 (2002), 583–639.

[25] A. Zahran, C. M. Anderson-Cook and R. H. Myers, Fraction of Design Space to Assess Prediction Capability of Response Surface Designs, J. Qual. Tech. 35 (2003), 377–386.