Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SIMULATION SELECTION FOR EMPIRICAL MODEL COMPARISON

Qiong Zhang
Yongjia Song

Department of Statistical Sciences and Operations Research
Virginia Commonwealth University
1015 Floyd Ave
Richmond, VA 23284, USA

ABSTRACT

We propose an efficient statistical method for the empirical model comparison, which is typically referred to as a simulation procedure to evaluate multiple statistical learning algorithms. First, we use experimental designs to appropriately construct the training and test sets for estimating the empirical performances of these models using the mean square errors. Second, we apply the idea of Bayesian fully sequential ranking and selection to optimally allocate the simulation budget according to the value of information. To make the procedure computationally tractable, we assume a normal-Wishart prior distribution, and propose a new approximation scheme for the posterior distribution by matching it with a normal-Wishart distribution using the first-order moment. Numerical experiments are conducted to show the superiority of the proposed approach on empirical model comparison problems.

1 INTRODUCTION

Empirical model comparison is typically referred to as a simulation procedure to evaluate multiple statistical learning algorithms, see for example, Caruana and Niculescu-Mizil (2006). In this procedure, statistical learning algorithms are applied to fit a large number of observational data sets of different types (e.g., complete or missing, large or small, etc.), and their predictive performances are evaluated through criteria such as area under curve (for binary response), mean square error (for continuous response), etc. The goal is to select the best algorithm in terms of the predictive performances.

In addition to model evaluation based on observational data sets, another field where the empirical model comparison procedure is widely applied is computer experiments (Sacks et al. 1989). Computer experiments refers to a number of runs of computer codes based on mathematical models for real systems, e.g., climate change, automobile crash, etc. Figure 1 illustrates the idea of empirical model comparison in the computer experiment setting. Instead of observational data, the input data sets for computer experiments are usually pre-specified using certain statistical designs. The outputs are obtained by running computer codes based on a complex mathematical model. A data set in computer experiments consists of the pre-specified input points and their corresponding outputs. A surrogate statistical model (Tai et al. 2006) is then built for this data set. Choosing an appropriate surrogate model is critical for computer experiments, and the choice is often made by empirically comparing the predictive performances of a set of candidate models.

In computer experiments, the computational effort is a fundamental concern. Running computer experiments could be extremely time consuming, so the data points available for empirically comparing surrogate models are usually limited, which leads to high variability in the empirical comparison results. To stabilize these comparison results, the predictive performances of surrogate models are usually evaluated by replicating the training and test procedure, using, e.g., multi-fold cross-validation, leave-one-out cross-validation, bootstrapping, etc. (Tai et al. 2006). However, these replications over a large number of surrogate

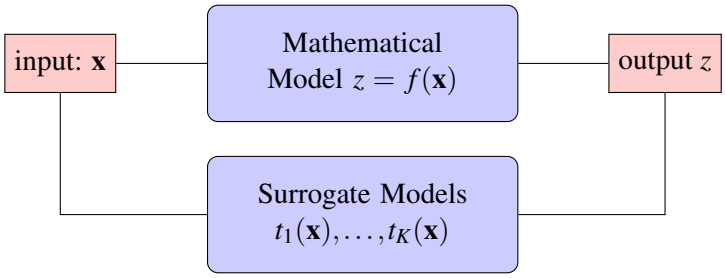

Figure 1: An illustration of computer experiments.

models could also be computationally expensive. Therefore, controlling the variability of comparison results under limited computational budget is a critical issue for computer experiments.

To address this issue, we focus on two strategies to balance the variability and the computational costs. First, we aim to reduce the variability of the estimated prediction errors by experimental design-based data splitting for training and test (Zhang and Qian 2013). Second, we aim to avoid unnecessary replications in empirical model comparison. For example, as shown in Figure 2, it is clear that the predictive performances of models 1-4 are significantly better than models 5-9. Therefore, replications on models 5-9 are not necessary, and more replications should be conducted on models 1-4 to further reduce the variability in these models. This observation motivates us to apply the ranking and selection procedure (Powell and Ryzhov 2012).

Ranking and selection tackles the problem of choosing a single alternative or multiple alternatives from a candidate set based on their random performances. The idea of ranking and selection is to replicate more on "promising" candidates. In the literature, ranking and selection is studied under two different streams. In the frequentists' perspective, ranking and selection is based on the indifferent-zone approach (Kim and Nelson 2001, Kim and Nelson 2006), see, e.g., Hong and Nelson (2009) or Kim and Nelson (2007) for a more extensive survey. This procedure has recently been applied in a parallel computing setting (Luo et al. 2014). From the Bayesian perspective, ranking and selection is studied under the "value of information" approach, see, e.g., Chick (2006), Powell and Ryzhov (2012) for overviews of the framework. Recently, the sequential selection procedures based on optimizing the expected opportunity cost according to the information obtained has been widely studied (Chick and Frazier 2012, Xie and Frazier 2013, Qu, Ryzhov, Fu, and Ding 2014).

In this paper, we study the empirical model comparison problem under a Bayesian framework, and develop an algorithm by customizing the fully sequential ranking and selection based on value of information (Powell and Ryzhov 2012). We focus on computer experiments with continuous outputs. In this case, the mean squared error (MSE) is used to evaluate the model performance by averaging the prediction errors of multiple test points. According to the Central Limit Theorem, the MSE approximately follows a normal distribution. However, due to the randomness of the input points, the correlation structure of the candidate models is unknown. Therefore, results in Powell and Ryzhov (2012) cannot be applied directly in our setting. For ranking and selection with unknown correlation structure, Qu et al. (2014) recently proposed an iterative procedure based on the normal-Wishart distribution (Gupta and Nagar 2000). To keep the procedure computationally tractable, the updating formula in Qu et al. (2014) is developed by approximating the posterior distribution as a normal-Wishart distribution via Kullback-Leibler divergence. In this paper, we propose a different approximation scheme to match the posterior distribution to a normal-Wishart distribution, using the idea of moments matching. Numerical results show the superiority of our approach on empirical model comparison problems.

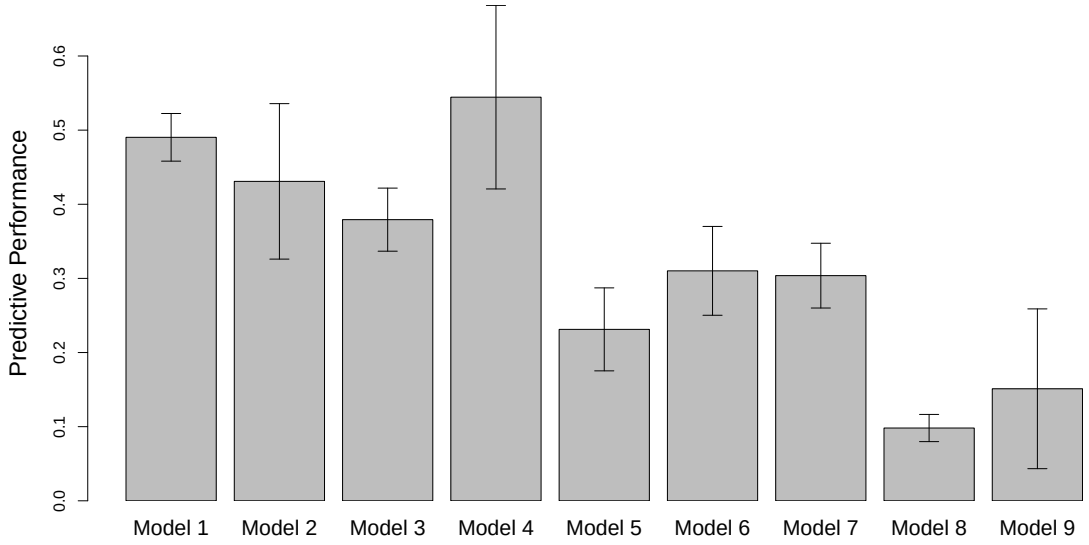

Figure 2: An illustration of model comparison according to their predictive performances (higher value means better performance). A 95% confidence interval of the predictive performance is provided for each model.

2 METHODOLOGY

2.1 The Setup for Empirical Model Comparison

We aim to select the best model among K surrogate models: $\{1,\ldots,K\}$ based on their predictive performances. For each model k, the predictive performance is estimated empirically using a data set $T = \{(\mathbf{x}_i, z_i)\}_{i=1}^{|T|}$, where $\{z_i\}_{i=1}^{|T|}$ are the outputs, and each $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional input vector. In this paper, the data set T is split into a training set D_{tr} and a test set D_{ts}. We use the MSE as the criterion to evaluate the predictive performance, that is, the model with a smaller MSE is considered to have a better predictive performance. For ranking and selection, the best alternative is usually assumed to be the one that achieves the largest quantity. To be consistent with the ranking and selection literature, we denote a sample from the k-th model in the ranking and selection procedure as the negative of the MSE:

$$\hat{y}_k = -\frac{1}{|D_{ts}|}\sum_{i\in D_{ts}}[z_i - t_k(x_i)]^2, \tag{1}$$

where $t_k(\cdot)$ is the predictor that is built based on the training data D_{tr}. We further denote $\hat{\mathbf{Y}}$ as a vector containing $(\hat{y}_1,\ldots,\hat{y}_K)^\top$. Given the data set D_{tr}, according to the Central Limit Theorem, when $|D_{ts}|$ is large enough:

$$\hat{\mathbf{Y}}|D_{tr} \to \mathcal{N}_k(\mu, \mathbf{R}), \tag{2}$$

where

$$\mu = \mathbb{E}_{D_{ts}}(\hat{\mathbf{Y}}|D_{tr}) \tag{3}$$

is the conditional mean vector of $\hat{\mathbf{Y}}$, and

$$\mathbf{R} = \left[\mathrm{var}_{D_{ts}}(\hat{\mathbf{Y}}|D_{tr})\right]^{-1} \tag{4}$$

is the precision matrix of $\hat{\mathbf{Y}}$. Since the conditional distribution of $\hat{\mathbf{Y}}|D_{tr}$ is determined by μ and $\mathbf{R}$, (2) can be written as:

$$\hat{\mathbf{Y}}|\mu,\mathbf{R} \rightarrow \mathscr{N}_k(\mu,\mathbf{R}). \tag{5}$$

The variability of the training data set D_{tr} is carried by μ and $\mathbf{R}$. We assume that μ and $\mathbf{R}$ follow a normal-Wishart prior distribution:

$$\mu|\mathbf{R} \sim \mathscr{N}_K(\theta^0, q^0\mathbf{R}),\ \mathbf{R} \sim \mathscr{W}_K(b^0, \mathbf{B}^0), \tag{6}$$

where given $\mathbf{R}$, the conditional distribution of μ is a multivariate normal distribution with mean vector θ^0 and precision matrix $q^0\mathbf{R}$, and $\mathbf{R}$ follows a Wishart distribution with parameters b^0 and $\mathbf{B}^0$. The conjugacy property of the normal-Wishart distribution allows us to update the distribution after each new observation in a computationally tractable way (DeGroot 2004).

2.2 The Normal-Wishart Updating Procedure by Replicating All K Models in Each Iteration

We first consider the case when the computational budget allows us to replicate all K models in each iteration. Assume that the setup in (5) and (6) holds for all the previous n iterations, and let the parameters of the prior distribution in the n-th iteration be q^n, b^n, θ^n, and $\mathbf{B}^n$. In the $(n+1)$-th iteration, we generate data set $T^{(n+1)}$ and split it into a training data set $D_{tr}^{(n+1)}$ and a test data set $D_{ts}^{(n+1)}$ (If a new data set $T^{(n+1)}$ is not available, the training data set $D_{tr}^{(n+1)}$ and the test data set $D_{ts}^{(n+1)}$ can be generated by randomly splitting a fixed T in each iteration). The MSE for all K models are calculated and collected in $\hat{\mathbf{Y}}^{(n+1)}$. Given the *entire* vector $\hat{\mathbf{Y}}^{n+1}$, the posterior distribution of μ and $\mathbf{R}$ in the $(n+1)$-th iteration is:

$$p^{n+1}(\mu,\mathbf{R}|\hat{\mathbf{Y}}^{n+1}) \propto p(\hat{\mathbf{Y}}^{n+1}|\mu,\mathbf{R})p^n(\mu|\mathbf{R})p^n(\mathbf{R}). \tag{7}$$

According to the normal-Wishart distribution,

$$p(\hat{\mathbf{Y}}^{n+1}|\mu,\mathbf{R}) \propto |\mathbf{R}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(\hat{\mathbf{Y}}^{n+1}-\mu)^\top\mathbf{R}(\hat{\mathbf{Y}}^{n+1}-\mu)\Big\}, \tag{8}$$

$$p^n(\mu|\mathbf{R}) \propto |\mathbf{R}|^{\frac{1}{2}} \exp\Big\{-\frac{q^n}{2}(\mu-\theta^n)^\top\mathbf{R}(\mu-\theta^n)\Big\}, \tag{9}$$

and

$$p^n(\mathbf{R}) \propto |\mathbf{R}|^{\frac{b^n-K-1}{2}} \exp\Big\{-\frac{1}{2}tr(\mathbf{B}^n\mathbf{R})\Big\}. \tag{10}$$

As noted in Qu et al. (2014) and an earlier literature (DeGroot 2004), the normal-Wishart distribution has the nice property that the posterior distribution of μ and $\mathbf{R}$ in (7) is still a normal-Wishart distribution. Thus, parameters in the $(n+1)$-th iteration can be updated by:

$$q^{n+1} = q^n + 1 \tag{11a}$$

$$b^{n+1} = b^n + 1 \tag{11b}$$

$$\theta^{n+1} = \frac{q^n\theta^n + \hat{\mathbf{Y}}^{n+1}}{q^n+1} \tag{11c}$$

$$\mathbf{B}^{n+1} = \mathbf{B}^n + \frac{q^n}{q^n+1}(\theta^n-\hat{\mathbf{Y}}^{n+1})(\theta^n-\hat{\mathbf{Y}}^{n+1})^\top. \tag{11d}$$

2.3 A Bayesian Updating Procedure Based on Moments Matching

The normal-Wishart distribution provides a very convenient way of updating our beliefs about the alternatives based on simple statistical estimation. However, this convenient way of updating requires the complete observation of $\hat{\mathbf{Y}}^{n+1}$ at each iteration, which could be time-consuming when the number of alternatives under comparison is large. Motivated by the fully sequential Bayesian ranking and selection procedure (Powell and Ryzhov 2012, Frazier, Powell, and Dayanik 2009), we may just choose the most promising alternative to simulate at each iteration. However, the flexibility of choosing only one alternative at a time will cause a significant challenge: the updating formula (11) cannot be applied in this case (Qu et al. 2014). Indeed, suppose in iteration $n+1$ we only compute the MSE of the k-th alternative, $\hat{y}_k^{n+1}$. It follows a conditional distribution $\mathcal{N}(\mu_k, (\mathbf{R}^{-1})_{kk}^{-1})$, where μ_k is the k-th element of μ, and $(\mathbf{R}^{-1})_{kk}$ is the k-th diagonal element of $\mathbf{R}^{-1}$. Using Bayes' rule, the posterior distribution of μ and $\mathbf{R}$ given $\hat{y}_k^{n+1}$ is:

$$\begin{aligned} p^{n+1}(\mu, \mathbf{R}|\hat{y}_k^{n+1}) \propto & |\mathbf{R}|^{\frac{b^n-K-1}{2}} \exp\Big\{-\frac{1}{2}tr(\mathbf{B}^n\mathbf{R})\Big\}|\mathbf{R}|^{\frac{1}{2}} \exp\Big\{-\frac{q^n}{2}(\mu-\theta^n)^\top \mathbf{R}(\mu-\theta^n)\Big\} \\ & \cdot (\mathbf{R}^{-1})_{kk}^{-1/2} \exp\Big\{-\frac{(\hat{y}_k^{n+1}-\mu_k)^2}{2(\mathbf{R}^{-1})_{kk}}\Big\}. \end{aligned} \tag{12}$$

From (12), we see that the posterior distribution is no longer a normal-Wishart distribution. Although the conditional distribution of μ given $\mathbf{R}$ is still a multivariate normal distribution, the marginal distribution of $\mathbf{R}$ is not a Wishart distribution, and therefore, the conjugacy property of normal-Wishart distribution cannot be applied.

To develop an updating formula based on (12), Qu et al. (2014) created an optimal approximation of conjugacy based on the Kullback-Leibler divergence between the posterior distribution (12) and a normal-Wishart distribution. This makes their learning model still computationally tractable, allowing their updating formula to be embedded into the fully sequential ranking and selection framework with a known correlation structure in Frazier, Powell, and Dayanik (2009). However, there exists a challenge about determining the step size for b^n in the updating procedure of Qu et al. (2014). Besides, matching two distributions through Kullback-Leibler divergence is a very strong condition. The Kullback-Leibler divergence of two distributions is equivalent to a distance measure of two distributions over the moments of all possible orders. In contrast, the parameters required in the updating formula only involve the first order moments. Therefore, a complete matching of two distributions may not be necessary. In our approach, we only match the first order moments between the posterior distribution (12) and a normal-Wishart distribution. This also makes the learning model computationally tractable, and there is no issue on determining the step size for b^n.

Based on (12), we provide the conditional expectation and conditional variance of μ given $\mathbf{R}$ and $\hat{y}_k^{n+1}$ in Proposition 1. Before formally presenting the result, we first introduce the notation here:

- For any vector $\mathbf{a}$ of size K, $\mathbf{a}_k$ is the k-th element of $\mathbf{a}$, and $\mathbf{a}_{-k} \in \mathbb{R}^{K-1}$ is the vector consisting of all the components of $\mathbf{a}$ except $\mathbf{a}_k$.
- For any $K \times K$ symmetric matrix $\mathbf{B}$, $\mathbf{B}_{kk}$ is the k-th diagonal element of $\mathbf{B}$, $\mathbf{B}_{\cdot,k}$ is the k-th column of $\mathbf{B}$, $\mathbf{B}_{-k,k} \in \mathbb{R}^{K-1}$ is the vector made by excluding the k-th element of $\mathbf{B}_{\cdot,k}$, $\mathbf{B}_{-k,-k}$ is the submatrix of $\mathbf{B}$ that corresponds to all but the k-th component, and also,

$$\mathbf{B}_{-k|k} := \mathbf{B}_{-k,-k} - \frac{\mathbf{B}_{-k,k}\mathbf{B}_{k,-k}}{\mathbf{B}_{kk}}.$$

Proposition 1 (a) Given $\mathbf{R}$ and $\hat{y}_k^{n+1}$, μ follows a multivariate normal distribution with mean

$$\mathbb{E}[\mu|\mathbf{R}, \hat{y}_k^{n+1}] = \theta^n + \frac{1}{(\mathbf{R}^{-1})_{kk}}(\mathbf{R}^{-1})_{\cdot,k}\frac{\hat{y}_k^{n+1}-\theta_k^n}{q^n+1}, \tag{13}$$

and variance matrix $\tilde{\Sigma}_{R,\hat{y}_k^{n+1}}$, whose elements are given by:

$$(\tilde{\Sigma}_{R,\hat{y}_k^{n+1}})_{kk} = (q^n+1)^{-1}(\mathbf{R}^{-1})_{kk}$$

$$(\tilde{\Sigma}_{R,\hat{y}_k^{n+1}})_{k,-k} = (q^n+1)^{-1}(\mathbf{R}^{-1})_{k,-k}$$

and

$$(\tilde{\Sigma}_{R,\hat{y}_k^{n+1}})_{-k,-k} = \frac{(\mathbf{R}^{-1})_{-k|k}}{q^n} + \frac{(\mathbf{R}^{-1})_{-k,k}(\mathbf{R}^{-1})_{k,-k}}{(q^n+1)(\mathbf{R}^{-1})_{kk}}.$$

(b) Let $\mathbb{E}_{\mathbf{R}}$ be the expectation taken with respect to $\mathbf{R}$.

$$\mathbb{E}_{\mathbf{R}}\left(\tilde{\Sigma}_{R,\hat{y}_k^{n+1}}\middle|\hat{y}_k^{n+1}\right) = \frac{1}{b^n-3K-2}\tilde{\mathbf{B}}^n \tag{14}$$

where

$$\tilde{\mathbf{B}}^n_{kk} = \mathbf{B}^n_{kk} + \frac{q^n}{q^n+1}(\hat{y}_k^{n+1}-\theta_k^n)^2$$

$$\tilde{\mathbf{B}}^n_{k,-k} = (\tilde{\mathbf{B}}^n_{-k,k})^\top = \mathbf{B}^n_{k,-k} + \frac{q^n}{q^n+1}\frac{\mathbf{B}^n_{k,-k}}{\mathbf{B}^n_{kk}}(\hat{y}_k^{n+1}-\theta_k^n)^2,$$

and

$$\tilde{\mathbf{B}}^n_{-k,-k} = \frac{q^n+1}{q^n}\mathbf{B}^n_{-k|k}$$

$$+\left(1+\frac{q^n}{q^n+1}(\mathbf{B}^n)^{-1}_{kk}(\hat{y}_k^{n+1}-\hat{\theta}_k^n)^2\right)\left((b^n-3K-2)^{-1}\mathbf{B}^n_{-k|k}+(\mathbf{B}^n)^{-1}_{kk}\mathbf{B}^n_{-k,k}\mathbf{B}^n_{k,-k}\right).$$

Furthermore,

$$\mathbb{E}_{\mathbf{R}}\left((\mathbf{R}^{-1})^{-1}_{kk}(\mathbf{R}^{-1})_{\cdot,k}|\hat{y}_k^{n+1}\right) = (\mathbf{B}^n)^{-1}_{kk}\mathbf{B}^n_{\cdot,k}. \tag{15}$$

According to Proposition 1, we match $\mathbb{E}(\mu|\hat{y}_k^{n+1})$ and $\mathbb{E}\left[(q^{n+1})^{-1}\operatorname{var}(\mu|R,\hat{y}_k^{n+1})|\hat{y}_k^{n+1}\right]$ under the probability distribution in (12) with the corresponding moments of a normal-Wishart distribution. In Proposition 2, we give the updating formula according to this normal-Wishart distribution. To guarantee that $\mathbb{E}(\mu|\hat{y}_k^{n+1})$ and $\mathbb{E}\left[(q^{n+1})^{-1}\operatorname{var}(\mu|R,\hat{y}_k^{n+1})|\hat{y}_k^{n+1}\right]$ based on the normal-Wishart distribution of the $(n+1)$-th iteration are equal to those based on the probability distribution in (12), we propose the following updating formulas.

Proposition 2 Let $q^{n+1}=q^n+1$ and $b^{n+1}=b^n+1$, the updating formulas of θ^{n+1} and $\mathbf{B}^{n+1}$ based on moment matching are given by:

$$\theta^{n+1} = \theta^n + \frac{\mathbf{B}^n_{\cdot,k}}{\mathbf{B}^n_{kk}}\frac{\hat{y}_k^{n+1}-\theta_k^n}{q^n+1} \tag{16}$$

and

$$\mathbf{B}^{n+1} = \frac{b^n-K}{b^n-3K-2}\tilde{\mathbf{B}}^n. \tag{17}$$

Based on this new updating formula, we apply the "look ahead" strategy in Powell and Ryzhov (2012) and Qu et al. (2014) to select the best alternative to sample at each iteration. The idea is to calculate the expected value of information for each alternativfe k at each iteration n:

$$\mathscr{V}_k(S^n) = \mathbb{E}^n\left[\max_i \theta_i^{n+1}|k^n=k\right] - \max_i \theta_i^n, \tag{18}$$

where $\mathbb{E}^n$ is the conditional expectation given the posterior distribution at iteration n, and k^n denotes the alternative to measure at iteration n. The expected value of information can be calculated based on the predictive distribution of θ^{n+1}, which depends on the updating formula for θ^{n+1} given a new sample $\hat{y}_k^{n+1}$. We adapt this strategy in our approach using our updating formula in Proposition 2.

3 NUMERICAL EXPERIMENTS

We present numerical experiment results to demonstrate the effectiveness of the proposed approach. In our experiment, we compare the following four different settings:

I Proposed: At each iteration, the updating formula in Proposition 2 is used, and the alternative to simulate is chosen based on the value of information.
II Qu: The Bayesian fully sequential ranking and selection procedure based on the value of information proposed in Qu et al. (2014) is applied.
III Random: At each iteration, the updating formula in Proposition 2 is used, but the alternative to simulate is chosen randomly.
IV All: At each iteration, all alternatives are simulated, and the standard updating formula (11) is applied.

The performances of the four methods are compared using their corresponding opportunity costs at each iteration. As in Qu et al. (2014), the opportunity cost of method m in iteration n is defined by

$$C_n^m = \max_k \mu_k^n - \mu^s_{\arg\max_k \theta_k^n}, \tag{19}$$

where μ^n is a vector of K elements, each of which is calculated by averaging the first n samples.

3.1 Data Generated From Multivariate Normal with A Normal-Wishart Prior

We first consider an example where the samples are exactly generated from a multivariate normal distribution with a normal-Wishart prior distribution. The purpose of this example is to show the performances of the four methods, "Proposed", "Qu", "Random", and "All", when the distribution of the samples exactly matches our assumption. In this example, we compare $K = 30$ alternatives. For the k-th alternative, we generate the sample $\hat{y}_k^{(n+1)}$ from the marginal distribution of a multivariate normal with a normal-Wishart prior distribution. The parameters of the prior distribution are specified as 1) $b = p = 60$, 2) the k-th element of θ is $0.1 \times (k-1)/K$, and 3) matrix $\mathbf{B}$ has diagonal elements $\mathbf{B}_{kk} = k+1$ and off-diagonal elements $\mathbf{B}_{kh} = 0.3^{|k-h|}$. After generating 10 samples from each model, we estimate the sample mean and sample covariance matrix to obtain the parameters in the prior distribution. We apply the four methods for $N = 1000$ iterations, and we show the opportunity costs in each iteration in Figure 3.

From Figure 3, we can see that the opportunity cost for the proposed method is significantly smaller than "Qu" when the number of samples is small. As the number of samples gets larger, their differences in the opportunity costs become smaller, and eventually the two opportunity costs become the same. Comparing the four methods all together, we can see that both option "Qu" and "Proposed" perform much better than the random selection procedure, but not as good as option "All". However, option "All" is significantly more computationally expensive since all alternatives are simulated at each iteration.

3.2 Empirical Model Comparison with Borehole Example

Next we test the four methods, "Proposed", "Qu", "Random", and "All", for empirical model comparison on the borehole function (Morris, Mitchell, and Ylvisaker 1993), a widely used example for illustrating various methods in computer experiments. The function models the flow rate of water through a borehole in the following form:

$$z = \frac{2\pi x_1 (x_2 - x_3)}{\log(x_4/x_5)\left[1 + \frac{2x_6 x_1}{\log(x_4/x_5) x_5^2 x_7} + \frac{x_1}{x_8}\right]}, \tag{20}$$

where z is the response of the function.

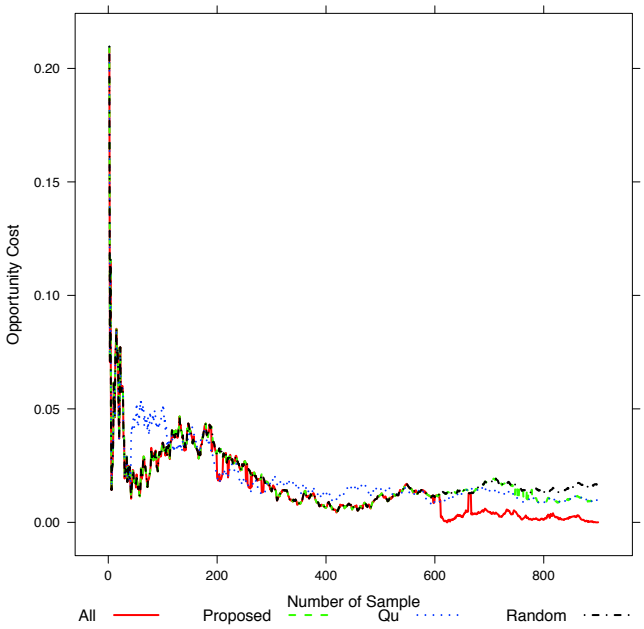

Figure 3: The opportunity cost comparison for the four settings when the data is generated from a multivariate normal distribution with a normal-Wishart prior.

The borehole function is well-known to be challenging to approximate. The mean squared errors given by the surrogate models are usually significantly larger than 0. Therefore, the sample mean squared error can be well approximated by a multivariate normal distribution when the test data set is large enough.

We generate the input points $\mathbf{x}_i's$ from a sliced Latin hypercube design (Qian 2012, Zhang and Qian 2013) with two slices and 100 data points. After calculating all the input-output pairs, we use the data set that corresponds to one slice of the Latin hypercube design as the training set, and use the rest of the data set as test set. We consider 15 Kriging predictors (Tai et al. 2006) using the training set, and compute their MSEs using the test set. These 15 Kriging predictors are constructed using different covariance functions and tuning parameters. To be more specific, they come from three covariance functions, exponential, stationary taper and wendland (see R package fields), and each covariance function is applied with five different tuning parameters: 0.1, 0.3, 0.5, 0.7, and 0.9. For the prior distribution, we run 10 samples initially, and use the sample mean and covariance as θ^0 and $\mathbf{B}^0$. We apply the four methods for $N = 200$ iterations. From iteration to iteration, we regenerate the sliced Latin hypercube design for the training data points and the test data points. So the randomness of the MSE comes from the randomness in these data points. We show the opportunity costs in each iteration in Figure 4.

From Figure 4, we can see that as the number of samples increases, the opportunity cost for the proposed method quickly decreases and becomes close to the opportunity cost for option "All". For options "Qu" and "Random", although their opportunity costs are smaller than "Proposed" when the number of samples is relatively small, when the sample size increases, their performances are outperformed by the proposed method. The computational results in this example show the superiority of the proposed method on empirical model validation over the naive method ("Random") and the state-of-the-art method by Qu et al. (2014).

We provide some insights on why option "Qu", which uses the approximation of the normal-Wishart distribution via the Kullback-Leibler divergence, is outperformed by the proposed method in the borehole example. First, the randomness of the training set is described by the Wishart distribution, and the randomness of the test set is described by the normal distribution. According to (1), when the test set is large enough, the

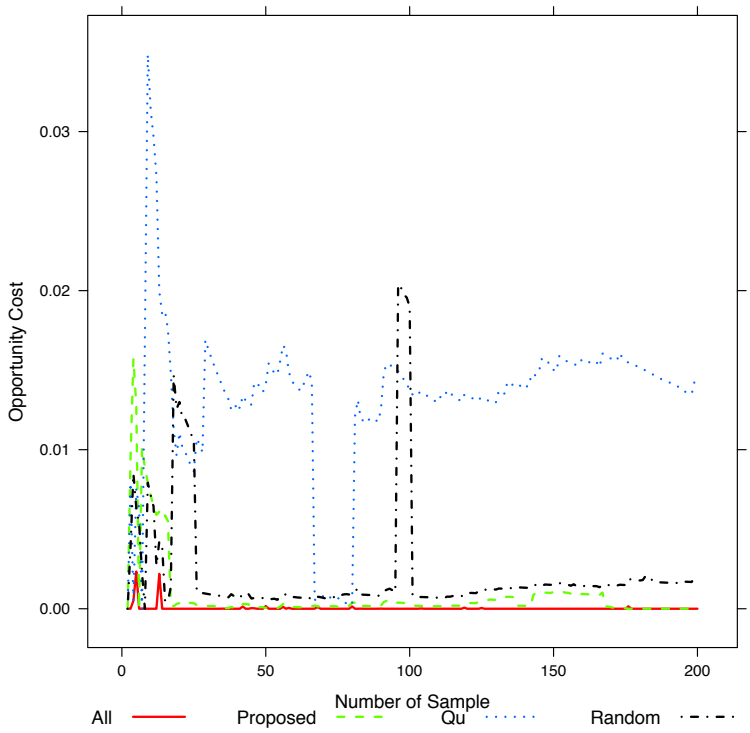

Figure 4: The opportunity costs for the four methods on their MSEs using the borehole function.

normal distribution in (6) is a good approximation. However, the Wishart distribution in (6) is only a rough approximation. Therefore, the more accurate approximation based on the Kullback-Leibler divergence may be overfitting compared to a less accurate approximation given by moment matching. Second, we observed in our numerical experiments that option "Qu" had some singularity issues when computing the precision matrices. This may explain the oscillation effect shown in Figure 4.

4 CONCLUDING REMARKS

We have presented an efficient statistical method for the empirical model comparison according to the predictive performance based on the MSE. We have applied the Bayesian fully sequential ranking and selection procedure based on the value of information, to optimally allocation the simulation effort for a set of candidate models. Following Qu et al. (2014), we have assumed a normal-Wishart prior distribution, and we have proposed a new approximation scheme by matching the posterior distribution with a normal-Wishart distribution using the first-order moments. Our preliminary numerical study has shown that our approximation scheme is effective on empirical model comparison examples in computer experiments. As an extension, this type of sequential learning strategy can also be applied to sequentially select design points in computer experiments.

A Proof of Proposition 1

Proof. First, we consider the distribution of μ given $\mathbf{R}$ and $\hat{y}_k^{n+1}$. Let μ^n be a random vector following a multivariate normal distribution with mean θ^n and precision matrix $q^n\mathbf{R}$. We further denote Y_k^n as a normal random variable with mean $\hat{y}_k^{n+1}$ and variance $(\mathbf{R}^{-1})_{kk}$. Assume that Y_k^n is independent with μ^n. Given $\mathbf{R}$

and $\hat{y}_k^{n+1}$,

$$\mu^{n+1} = \mu^n + \frac{\mathbf{R}^{-1}_{\cdot,k}}{\mathbf{R}^{-1}_{kk}} \frac{Y_k^n - \theta_k^n}{q^n + 1} \tag{21}$$

is a random vector whose the probability distribution is proportional to (12). Since μ^{n+1} is a linear combination of normal random vector μ^n and Y_k^n, it still follows a multivariate normal distribution, whose mean and variance can be easily derived as in (a).

We now prove (b). The conditional distribution of $\mathbf{R}$ given $\hat{y}_k^{n+1}$ is

$$\begin{aligned}
p(\mathbf{R}|\hat{y}_k^{n+1}) &= \int p(\mu, \mathbf{R}|\hat{y}_k^{n+1}) d\mu = \int\int p(\mu, \mathbf{R}|\hat{y}_k^{n+1}) d\mu_{-k} d\mu_k \\
&\propto |\mathbf{R}|^{\frac{b^n - K}{2}} (\mathbf{R}^{-1})_{kk}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^n \mathbf{R})\right\} \\
&\int \exp\left\{-\frac{(\hat{y}_k^{n+1} - \mu_k)^2}{2(\mathbf{R}^{-1})_{kk}}\right\} d\mu_k \int \exp\left\{-\frac{q^n}{2}(\mu - \theta^n)^\top R(\mu - \theta^n)\right\} d\mu_{-k} \\
&\propto |\mathbf{R}|^{\frac{b^n - K - 1}{2}} (\mathbf{R}^{-1})_{kk}^{-1} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^n \mathbf{R})\right\} \int \exp\left\{-\frac{(\hat{y}_k^{n+1} - \mu_k)^2}{2(\mathbf{R}^{-1})_{kk}} - \frac{q^n(\mu_k - \theta_k^n)^2}{2(\mathbf{R}^{-1})_{kk}}\right\} d\mu_k \\
&\propto |\mathbf{R}|^{\frac{b^n - K - 1}{2}} (\mathbf{R}^{-1})_{kk}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{B}^n \mathbf{R})\right\} \exp\left\{-\frac{q^n(\hat{y}_k^{n+1} - \theta_k^n)^2}{2(q^n+1)(\mathbf{R}^{-1})_{kk}}\right\}.
\end{aligned}$$

Since $\tilde{\Sigma}_{\mathbf{R}, \hat{y}_k^{n+1}}$ involves $\mathbf{R}^{-1}$, we try to represent $p(\mathbf{R}|\hat{y}_k^{n+1})$ in terms of $\mathbf{R}^{-1}$.

$$\begin{aligned}
p(\mathbf{R}|\hat{y}_k^{n+1}) \propto (\mathbf{R}^{-1})_{kk}^{-\frac{b^n - K}{2}} |(\mathbf{R}^{-1})_{-k|k}|^{-\frac{b^n - K - 1}{2}} \exp\Bigg\{ &\frac{-q^n(\hat{y}_k^{n+1} - \theta_k^n)^2}{2(q^n+1)(\mathbf{R}^{-1})_{kk}} - \frac{1}{2}\operatorname{tr}\left(\mathbf{B}_{-k,-k}\left[(\mathbf{R}^{-1})_{-k|k}\right]^{-1}\right) \\
&+ \frac{(\mathbf{R}^{-1})_{k,-k}\left[(\mathbf{R}^{-1})_{-k|k}\right]^{-1}}{(\mathbf{R}^{-1})_{kk}} \mathbf{B}_{-k,k} - \frac{1}{2}\mathbf{B}_{kk}\left[\frac{1}{(\mathbf{R}^{-1})_{kk}} + \frac{(\mathbf{R}^{-1})_{k,-k}\left[(\mathbf{R}^{-1})_{-k|k}\right]^{-1}(\mathbf{R}^{-1})_{-k,k}}{(\mathbf{R}^{-1}_{kk})^2}\right]\Bigg\}.
\end{aligned}$$

First, we are interested in $p\left((\mathbf{R}^{-1})_{-k,k}\middle|(\mathbf{R}^{-1})_{kk}, (\mathbf{R}^{-1})_{-k|k}, \hat{y}_k^{n+1}\right)$, so we put $(\mathbf{R}^{-1})_{-k,k}$ terms together in $p(\mathbf{R}|\hat{y}_k^{n+1})$, and obtain:

$$p\left((\mathbf{R}^{-1})_{-k,k}\middle|(\mathbf{R}^{-1})_{kk}, (\mathbf{R}^{-1})_{-k|k}, \hat{y}_k^{n+1}\right) \sim \mathcal{N}\left(\frac{\mathbf{B}_{-k,k}(\mathbf{R}^{-1})_{kk}}{\mathbf{B}_{kk}}, \frac{(\mathbf{R}^{-1})_{kk}^2}{\mathbf{B}_{kk}}(\mathbf{R}^{-1})_{-k|k}\right).$$

Now we calculate $p\left((\mathbf{R}^{-1})_{kk}, (\mathbf{R}^{-1})_{-k|k}\right) = \int p\left((\mathbf{R}^{-1})_{-k,k}, (\mathbf{R}^{-1})_{kk}, (\mathbf{R}^{-1})_{-k|k}\right) \mathrm{d}(\mathbf{R}^{-1})_{-k,k}$. Putting all irrelevant terms outside of the integration, and using the multivariate normal density function, we have:

$$\begin{aligned}
p\left((\mathbf{R}^{-1})_{kk}, (\mathbf{R}^{-1})_{-k|k}\right) \propto (\mathbf{R}^{-1})_{kk}^{\frac{3K - b^n - 2}{2}} \cdot \exp&\left\{\frac{1}{(\mathbf{R}^{-1})_{kk}}\left[\frac{q^n(\hat{y}_k^{n+1} - \theta_k^n)^2}{2(q^n+1)} - \frac{\mathbf{B}_{kk}}{2}\right]\right\} \\
&\cdot |(\mathbf{R}^{-1})_{-k|k}|^{\frac{K - b^n + 2}{2}} \cdot \exp\left\{\frac{1}{2}\left[\frac{\mathbf{B}_{-k,k}^\top \left((\mathbf{R}^{-1})_{-k|k}\right)^{-1} \mathbf{B}_{-k,k}}{\mathbf{B}_{kk}} - \operatorname{tr}\left(\mathbf{B}_{-k.-k}(\mathbf{R}^{-1})_{-k|k}\right)\right]\right\}.
\end{aligned}$$

This means that both $(\mathbf{R}^{-1})_{kk}$ and $(\mathbf{R}^{-1})_{-k,k}$ are inverse-Wishart distribution:

$$(\mathbf{R}^{-1})_{-k|k} \sim \mathscr{W}^{-1}\left(\mathbf{B}_{-k.-k} - \frac{\mathbf{B}_{-k,k}\mathbf{B}_{-k,k}^{\top}}{\mathbf{B}_{kk}}, b^n - 2K - 2\right)$$

$$(\mathbf{R}^{-1})_{kk} \sim \mathscr{W}^{-1}\left(\mathbf{B}_{kk} + \frac{q^n(\hat{y}_k^{n+1} - \theta_k^n)^2}{q^n + 1}, b^n - 3K\right).$$

From the properties of above distributions, we can see that (b) holds.

□

B Proof of Proposition 2

Proof. First, based on (15) in Proposition 1, (16) can be derived by taking expectation of the right hand side of (13) with regard to $\mathbf{R}$. Second, the update in (17) is derived to guarantee that

$$(b^{n+1} - K + 1)^{-1}\mathbf{B}^{n+1} = \mathbb{E}_{\mathbf{R}}\left\{(q^{n+1})^{-1}\,\mathrm{var}_{\mu}(\mu|\mathbf{R}, \hat{y}_k^{n+1})|\hat{y}_k^{n+1}\right\}.$$

□

REFERENCES

Caruana, R., and A. Niculescu-Mizil. 2006. "An empirical comparison of supervised learning algorithms". In *ICML '06 Proceedings of the 23rd international conference on Machine learning*, edited by J. H. K. P. W. M. F. S. Jain, R. R. Creasey, 161–168.

Chick, S. 2006. "Bayesian ideas and discrete event simulation: why, what and how". In *Proceedings of the Winter Simulation Conference*, edited by L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto, 96–105.

Chick, S., and P. Frazier. 2012. "Sequential Sampling for Selection with Economics of Selection Procedures". *Management Science* 58 (3): 550–569.

DeGroot, M. 2004. *Optimal statistical decisions*. John Wiley and Sons.

Frazier, P., W. Powell, and S. Dayanik. 2009. "The knowledge-gradient policy for correlated normal rewards". *INFORMS Journal on Computing* 21 (4): 599–613.

Gupta, A., and D. Nagar. 2000. *Matrix variate distributions*. Chapman & Hall.

Hong, L., and B. Nelson. 2009. "A brief introduction to optimization via simulation". In *Proceedings of the Winter Simulation Conference*, edited by M. Rosetti, R. Hill, B. Johansson, A. Dunkin, and R. Ingalls, 75–85.

Kim, S.-H., and B. Nelson. 2001. "A fully sequential procedure for indifference-zone selection in simulation". *ACM Transactions on Modeling and Computer Simulation* 11 (3): 251–273.

Kim, S.-H., and B. Nelson. 2006. "On the asymptotic validity of fully sequential selection procedures for steady-state simulation". *Operations Research* 54 (3): 475–488.

Kim, S.-H., and B. Nelson. 2007. "Recent advances in ranking and selection". In *Proceedings of the Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 162–172.

Luo, J., L. Hong, B. Nelson, and Y. Wu. 2014. "Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environment". Submitted for publication.

Morris, M., T. Mitchell, and D. Ylvisaker. 1993. "Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction". *Technometrics* 35:243–255.

Powell, W., and I. Ryzhov. 2012. *Optimal learning*. John Wiley and Sons.

Qian, P. Z. G. 2012. "Sliced Latin hypercube designs". *Journal of the American Statistical Association* 107 (497): 393–399.

Qu, H., I. Ryzhov, M. Fu, and Z. Ding. 2014. "Sequential selection with unknown correlation structures". Submitted for publication.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. "Design and analysis of computer experiments". *Statistical Science* 4:409–435.

Tai, F. K., R. Li, and A. Sudjianto. 2006. *Design and modeling for computer experiments*. Boca Raton, FL: Chapman & HallCRC.

Xie, J., and P. Frazier. 2013. "Sequential Bayes-Optimal Policies for Multiple Comparisons with a Known Standard". *Operations Research* 61 (3): 1174–1189.

Zhang, Q., and P. Z. G. Qian. 2013. "Designs for crossvalidating approximation models". *Biometrika* 100:997–1004.

AUTHOR BIOGRAPHIES

Qiong Zhang is an Assistant Professor of statistics at Virginia Commonwealth University, Richmond, VA. She holds Ph.D. degree in statistics from University of Wisconsin-Madison. Her research interests include computer experiments, uncertainty quantification and spatial and spatial-temporal modeling. She is a member of ASA and INFORMS. Her email address is qzhang4@vcu.edu.

Yongjia Song is an Assistant Professor of operations research at Virginia Commonwealth University, Richmond, VA. He holds a Ph.D. in industrial engineering from University of Wisconsin-Madison. His research interests include optimization under uncertainty, integer programming, and applications of optimization. He is a member of INFORMS, SIAM, and MOS. His e-mail address is ysong3@vcu.edu.