A BAYESIAN METAMODELING APPROACH FOR STOCHASTIC SIMULATIONS

Jun Yin
Szu Hui Ng
Kien Ming Ng

Department of Industrial & Systems Engineering
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260, SINGAPORE

ABSTRACT

In the application of kriging model in the field of simulation, the parameters of the model are likely to be estimated from the simulated data. This introduces parameter estimation uncertainties into the overall prediction error, and this uncertainty can be further aggravated by random noise in the stochastic simulation. In this paper, a Bayesian metamodeling approach for kriging prediction is proposed for stochastic simulations to more appropriately account for the parameter uncertainties. The approach is first illustrated analytically using a simplified two point example. A more general Markov Chain Monte Carlo analysis approach is subsequently proposed to handle more general assumptions on the parameters and design. The general MCMC approach is compared with the modified nugget effect kriging model based on the M/M/1 simulation system. Initial results indicate that the Bayesian approach has better coverage and closer predictive variance to the empirical value than the modified nugget effect kriging model, especially in the cases where the stochastic variability is high.

1   INTRODUCTION

Discrete event simulation is a powerful tool in understanding and evaluating the performance of complex systems. Simulation models are applied in many disciplines, including manufacturing, epidemiology and ecology. Analysts use simulation models because costs, time or other constraints prohibit experimentation with real systems. In practice however, a single simulation run of a complex model may require a substantial use of computer resources and time. Moreover, these simulation models themselves may be quite complex so simpler approximations known as metamodels are often constructed. These metamodels attempt to accurately capture the relationships between the inputs and outputs of the simulation models, and are computationally more efficient.

The most popular technique used for metamodeling has been based on parametric polynomial response surface approximations. Although polynomial response metamodels offer good approximations for simple models, the main drawback of polynomial metamodels is their lack of flexibility to achieve a global fit. Furthermore, many of the complex simulation models and non stationary models have highly nonlinear responses so linear polynomial approximations are not adequate. Various other types of metamodels like multivariate adaptive regression splines, kriging, radial basis functions, artificial neural networks, and support vector regression (SVR), have been proposed in recent years. A review of these metamodel performances and applications in engineering and decision support systems can be found in (Simpson et al. 2001), (Li et al. 2010).

Among the different types of metamodels, the kriging model is one of the more promising metamodels as it is more adaptable than the regression based models and not as complicated and time consuming as artificial intelligence techniques. Although kriging originated in geostatistics (Matheron 1963), it has been successfully applied in many deterministic computer experiments ((Pham and Wagner 1994), (Roshan 2006), (Gupta et al. 2006), (Wu and Sun 2007)). For these deterministic simulation problems, the kriging model's interpolating characteristic is appropriate and attractive, providing predictions with the same values as the observations. Recently, there has also been an increasing interest in

adopting kriging metamodels for stochastic simulations, including discrete event simulation. Under homoscedastic assumptions on the noise, the kriging model with the nugget effect, which was originally proposed by (Cressie 1993) to account for measurement error in spatial statistics, has been applied by (Huang et al. 2006). The application of this model to stochastic simulation has been very successful if the underlying homoscedastic assumptions are met. However, the performance deteriorates quickly when the noise varies ((Yin, Ng, and Ng 2008), (Li et al. 2010)). (Kleijnen and Beers 2005) proposed two methods to improve the application of kriging in stochastic problems: the replication method and the studentization method. Both methods essentially converts the general heteroscedastic problem into a homoscedastic problem and applying the traditional kriging or nugget effect model. These methods however require sufficient computing resources and prior information about the variance function. (Yin, Ng, and Ng 2008) and (Ankenman, Nelson, and Staum 2010) propose the modified nugget effect model and the stochastic kriging model respectively to address the more general heteroscedastic case. Although these models enable much better modeling of stochastic systems, several issues in parameter estimation arise when the noise levels are high (Yin, Ng, and Ng 2009).

In this paper, we propose a Bayesian approach for kriging metamodeling for stochastic simulations. This general approach overcomes some of the problems identified in (Yin, Ng, and Ng 2009) by appropriately accounting for the parameter uncertainties in the predictor.

This paper proposes a general kriging metamodel for stochastic simulations, of which existing work of (Yin, Ng, and Ng 2008) is a special case. In this model, uncertainty in the regression parameters and the noise levels are accounted for, and we discuss a general approach to fully account for all parameter estimation and uncertainties under different assumptions over several given special cases and examples. This paper is organized as follows. In Section 2, we first describe our kriging metamodel formulation, then derive the predictive distributions of the model for specific special cases. Then, we propose a Markov Chain Monte Carlo (MCMC) approach to derive the predictive distribution for the general case. Finally in section 3, a comparison between the modified nugget effect model and the proposed Bayesian approach for kriging metamodeling is provided based on the M/M/1 queue.

2 MODELING APPROACH

Let $Y^s(x)$ be the output from the stochastic simulation at x_i, where $x_i = (x_1, \ldots, x_p)_i^T$ is a point in a p-dimensional input space. We assume that $Y^s(x_i)$ can be described by

$$Y^s(x_i) = Z(x_i) + \xi(x_i) \quad i = 1, \ldots, n \tag{1}$$

where $Z(x)$ is a *Gaussian process* with mean $\mu(x) = F^T(x)\beta$ and covariance function $\sigma_z^2 R_z$, which we denote as $GP(\mu(x), \sigma_z^2 R_z)$, and $\xi(x)$ is also a Gaussian process with with mean 0 and covariance function $\sigma_m^2 + \sigma_e^2 R_e$, $GP(0, \sigma_m^2 I + \sigma_e^2 R_e)$. $F(x)$ is a vector of q functions of x, and I is the identity matrix. We further assume that Z and ξ are independent. This general modeling of ξ as a *Gaussian process* enables a wider class of stochastic models to be modeled. σ_m^2 can be typically thought of as the white noise or base measurement error inherent throughout the system and $\sigma_e^2 R_e$ as the additional variability that is dependent on the input location. Describing $\xi(\cdot)$ in this manner enables the error covariances to be modeled as functions of locations and can also provide estimation of the noise at unobserved points by "borrowing" information on variability from other nearby points. This formalizes the ideas in (den Hertog, Kleijnen, and Siem 2005) and (Yin, Ng, and Ng 2008) for estimating variances at unobserved points where ad hoc interpolation and bootstrapping techniques were taken. The general form of (1) is similar to the form proposed in (Ankenman, Nelson, and Staum 2010) when the defined intrinsic variance is estimated. When R_e is diagonal, (1) reduces to the independent noise model adopted by (Yin, Ng, and Ng 2008) and (Ankenman, Nelson, and Staum 2010).

The correlation function R can have different forms, and usually depends on correlation parameters ϕ. A popular choice is the power exponential family of correlation functions which is commonly applied in the field of computer experiment for its smooth response characteristics,

$$R_z(x_i, x_j) = \prod_{k=1}^{p} \exp(-\phi_{z,k}(x_{i,k} - x_{j,k})^q), R_e(x_i, x_j) = \prod_{k=1}^{p} \exp(-\phi_{e,k}(x_{i,k} - x_{j,k})^q) \tag{2}$$

The scale correlation parameters $\phi_{z,k}$ and $\phi_{e,k}$ are the sensitivity parameters that control how fast the correlation decays with distance in the kth dimension, and the parameter q controls in general the

1056

smoothness of the response. When $q = 1$, (2) is commonly known as the exponential correlation function, and when $q = 2$, it is known as the Gaussian correlation function. Other forms of correlation functions including the linear correlation function given in the Eq. (3) below, the Matérn class of functions, the rational quadratic class of functions and the spline functions (more details can be found in Cressie (1993)).

$$R_z(x_i.x_j) = \prod_{k=1}^{p} 1 - \phi_{z,k}|x_{i,k} - x_{j,k}|, R_e(x_i,x_j) = \prod_{k=1}^{p} 1 - \phi_{e,k}|x_{i,k} - x_{j,k}| \tag{3}$$

2.1 Modeling Uncertainty

In the model defined in (1) and the correlation function in (2) or (3), the parameters $\theta = (\beta, \sigma_z^2, \phi_z, \sigma_e^2, \sigma_m^2, \phi_e)$ are typically unknown. One Bayesian approach to quantify this uncertainty is to assign a prior distribution $p(\beta, \sigma_z^2, \phi_z, \sigma_e^2, \sigma_m^2, \phi_e)$ on the parameters.

In this paper, we assume that $\beta, \sigma_z^2, \phi_z$ and $\sigma_e^2, \sigma_m^2, \phi_e$ are independent. We also assume that β, σ_z^2 and ϕ_z are independent, and σ_e^2 and σ_m^2 and ϕ_e are independent. Then, the prior on the parameters can be written as

$$\begin{aligned} p(\theta) = p(\beta, \sigma_z^2, \phi_z, \sigma_e^2, \sigma_m^2, \phi_e) &= p(\beta, \sigma_z^2)p(\phi_z)p(\sigma_e^2)p(\sigma_m^2)p(\phi_e) \\ &= p(\beta|\sigma_z^2)p(\sigma_z^2)p(\phi_z)p(\sigma_e^2)p(\sigma_m^2)p(\phi_e). \end{aligned} \tag{4}$$

This modeling approach also facilitates the accounting of any priori knowledge on the parameters to be accounted for. In the case with little prior information, the non informative priors can be used.

2.2 Bayesian Prediction and Predictive Distribution

Based on (1) and (4), to predict the response $Y^s(x_0) = Z(x_0)$ at any x_0, we first obtain the predictive distribution of $Y^s(x_0)$. The posterior predictive distribution at x_0 can be obtained by

$$f(Z(x_0)|Y_n) = \int_{\theta} f(Z(x_0)|Y_n, \theta) \times f(\theta|Y_n)d\theta. \tag{5}$$

The mean of this predictive distribution $\mu_{0|n}(x_0)$ is the best MSPE predictor at x_0, and variance can be interpreted as the mean squared prediction error of $\mu_{0|n}(x_0)$ (Santner, Williams, and Notz 2003). The advantage of this Bayesian approach is that it accounts for the parameter uncertainties, and hence can avoid the bias introduced with point estimates as described in (Zimmerman and Cressie 1992) and (Yin, Ng, and Ng 2009).

2.2.1 Derivation of the Predictive Distribution (*Assuming R and r_0 are known*)

In this section, to ease the computational burden, we first develop a prediction procedure by assuming that the parameters in the correlation function R and r_0 are known, i.e. ϕ_z and ϕ_e are known. We adopt an empirical Bayes approach and discuss later how these parameters can be estimated.

With ϕ_z and ϕ_e known and rewriting the remaining parameters $\theta_1 = (\beta, \sigma_z, \sigma_e, \sigma_m)$, we have from (5),

$$\begin{aligned} f(Z(x_0)|Y_n) &= \iiint f(Z(x_0)|Y_n, \beta, \sigma_z^2, \sigma_e^2, \sigma_m^2)f(\beta, \sigma_z^2, \sigma_e^2, \sigma_m^2)d\beta d\sigma_z^2 d\sigma_e^2 d\sigma_m^2 \\ &= \iiint f(Z(x_0)|Y_n, \beta, \sigma_z^2, \sigma_e^2, \sigma_m^2)f(\beta|\sigma_z^2, Y_n)f(\sigma_z^2, \sigma_e^2, \sigma_m^2|Y_n)d\beta d\sigma_z^2 d\sigma_e^2 d\sigma_m^2 \end{aligned} \tag{6}$$

A simple two point problem A simple two point problem is used to illustrate this approach. We use the following simplified two point problem to illustrate the approach analytically and to provide some insights.

Suppose two observations y_1 and y_2 are obtained at points x_1 and x_2 respectively, where the noise variability at each point is the same at $(\sigma_m^2 + \sigma_e^2)$. The distance between these two points
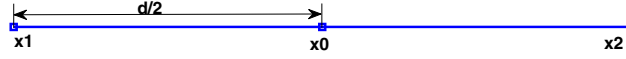
Figure 1: Design of the two point problem.

is d, and the prediction point x_0 is located at the center of the two points. We further assume that the correlation function between $Z(x_1)$ and $Z(x_2)$ is the linear correlation function in Eq. (3), given as $\rho = 1 - \phi_z d$. For ease of the integration, we introduce a new parameter τ defined as $\tau = (\sigma_m^2 + \sigma_e^2)/\sigma_z^2$, indicating loosely the ratio of the error noise to the signal function variation. For computational simplicity, we adopt non informative priors for β, σ_z^2 and τ with the following forms: $p(\beta) = 1, p(\sigma_z^2) = 1/\sigma_z^2, p(\tau) = 1/(2(1+\tau)^2)$. The priors for β and σ_z^2 are similar to the non informative priors proposed in (Santner, Williams, and Notz 2003) for Gaussian random function models of similar forms. The (Jeffreys 1961) prior is adopted for τ.

Then the covariance matrix $\Sigma = \sigma_z^2 R_z + \sigma_z^2 \tau R_e$, covariance vector $r(x_0)$, observation vector Y_n and the regressor for the observation points and prediction point are given respectively as:

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_z^2 \rho \\ \sigma_z^2 \rho & \sigma_z^2 \end{pmatrix} + \begin{pmatrix} \sigma_z^2 \tau & 0 \\ 0 & \sigma_z^2 \tau \end{pmatrix}, \tag{7}$$

$$r = (\frac{1}{2} + \frac{\rho}{2}, \frac{1}{2} + \frac{\rho}{2}),$$

$$Y_n = (y_1, y_2),$$

$$F(x) = (1,1), F(x_0) = 1.$$

The predictive distribution at x_0 can then be derived to be (Proof see Appendix C):

$$f(Z(x_0)|Y_n) \propto \sqrt{1 - \rho^2} (y_1^2 + y_2^2 - 2(y_1 + y_2)Z(x_0) + 2Z(x_0)^2)^{-1/2}.$$

Under these modeling assumptions, the predictive distribution at x_0 is a non-central t-distribution with mean $\frac{y_1 + y_2}{2}$ and variance $\frac{(y_1 - y_2)^2}{4}$. Using the mean as the predictor, the prediction of $Z(x_0)$ at point x_0 is given as $\frac{y_1 + y_2}{2}$ and the MSPE of this predictor is $\frac{(y_1 - y_2)^2}{4}$. With the linear correlation function and the non-informative priors, this result is intuitive as the mean predictor is simply the sample mean between the two observations, and the predictor variance is the sample covariance between these two points. The $100(1-\alpha)\%$ prediction interval is given as $\frac{y_1 + y_2}{2} \pm \frac{(y_1 - y_2)^2}{4} \cdot t_{0,\alpha/2}$.

A general prior setup In the remainder of this paper, to avoid improper posteriors, we adopt the following proper hierarchical prior setup

$$p(\beta|\sigma_z^2) \sim N_p(b_0, \sigma_z^2 V_z) \tag{8}$$

$$p(\sigma_z^2) \sim IG(\alpha_z, \gamma_z)$$

$$p(\sigma_e^2) \sim IG(\alpha_e, \gamma_e)$$

$$p(\sigma_m^2) \sim IG(\alpha_m, \gamma_m)$$

$$p(\phi_{z,i}) \sim G(a_z, b_z) \text{ for } i = 1, 2, \ldots, k$$

$$p(\phi_{e,i}) \sim G(a_e, b_e) \text{ for } i = 1, 2, \ldots, k$$

where $N_p(m,K)$ denotes a p-dimensional normal with mean m and covariance matrix K, $IG(\alpha,\gamma)$ denotes a Inverse Gamma distribution with density function,

$$p(s) = \frac{\gamma^\alpha}{\Gamma(\alpha)} s^{-(\alpha+1)} e^{-(\frac{\gamma}{s})}, s > 0, \alpha, \gamma > 0,$$

and $G(a,b)$ denotes the gamma distribution with density function

$$p(s) = \frac{b^a}{\Gamma(a)} s^{a-1} e^{-bs}, s > 0, a, b > 0.$$

The selection of the prior distributions follows the conjugate prior for the Gaussian Random Process in Gelman et al. (2003) and similar prior setting can be found in (Santner, Williams, and Notz 2003).

A further simplification when τ is known   Assuming τ is known (we discuss the estimation of τ in the next subsection), we obtain the following proposition.

Proposition 1:

Assuming that ϕ_z, ϕ_e and τ are known, the predictive distribution at x_0, is a non-central t distribution

$$Z(x_0)|y_n \sim T_1(\alpha_{z|y_n}, \mu(x_0)_{|y_n}, \sigma^2_{|y_n}(x_0)|y_n, \phi_z, \phi_e, \tau)$$

where

$$\alpha_{z|y_n} = n + 2\alpha_z,$$

$$\mu_{|y_n}(x_0) = F(x_0)\lambda/M + c_z^T(x_0)R^{-1}(y_n - F(x)\lambda/M),$$

and

$$\sigma^2_{|y_n}(x_0) = \frac{2\gamma_z + y_n^T R^{-1} y_n + b_0^T V_0^{-1} b_0 - \lambda^T \lambda/M}{\alpha_{z|y_n}}(1 - c_z^T(x_0)R^{-1}c_z(x_0) + h^T h/M),$$

with

$$\lambda = F^T R^{-1} Y_n + b_0 V_0^{-1}, M = F^T R^{-1} F + V_0^{-1}, h = r_z^T(x_0)R^{-1}F - F^T(x_0).$$

The proof of the Proposition 1 is given in the Appendix A.

Based on the above, a $100(1-\alpha)\%$ prediction interval for any point x_0 is given as $\mu(x_0)_{|y_n} \pm \sigma^2_{|y_n}(x_0) \cdot t_{\alpha_{z|y_n}/2, \alpha/2}$, where $t_{\alpha_{z|y_n}/2, \alpha/2}$ is a univariate non central t distribution with $\alpha_{z|y_n}/2$ degrees of freedom.

Estimation of parameters in R, r_0 and τ

The results from the above sections are based on the assumption that the parameters ϕ_z, ϕ_e, and τ are known. We adopt an empirical Bayes approach here by estimating these parameters for computational convenience. Here we propose an approach on how to estimate these parameters using the maximum likelihood method. This is similar to the approach taken by (Wang, Chen, and Tsui 2009). In Section 2.2, we consider a fully Bayesian approach which explicitly incorporates the uncertainties of these parameters in the modeling framework and propose a MCMC approach to derive the predictive distribution.

The maximum likelihood estimates for ϕ_z, ϕ_e, and τ are the parameter values that maximize the marginal likelihood $f(Y_n|\phi_z, \phi_e, \tau)$. Hence,

$$\{\hat{\phi}_z, \hat{\phi}_e, \hat{\tau}\} = \underset{\phi_z, \phi_e, \tau}{arg\max} f(y_n|\phi_z, \phi_e, \tau)$$

This can be solved using standard optimization functions in MATLAB.

An alternative approach to using the above maximum likelihood approach is to first estimate posterior distribution of the parameters ϕ_z, ϕ_e, and τ, and then using the posterior mean or mode of these distributions to estimate ϕ_z, ϕ_e, and τ. A similar approach to estimate these parameters is taken by Qian and Wu (2008).

2.2.2 A General Approach to Deriving the Predictive Distribution (when all the parameters are unknown)

In general, the integration in (5) can be done numerically. Assuming the prior setup in (8), full conditionals $f(\phi_z, \phi_e | Y_n, \sigma_z^2, \sigma_e^2, \sigma_m^2)$ and $f(\sigma_z^2, \sigma_e^2, \sigma_m^2 | Y_n, \phi_z, \phi_e)$ are obtainable (see Appendix B), and a Markov chain Monte Carlo approach can be applied in the following manner:

1   Set an initial starting value for $(\sigma_z^2, \sigma_e^2, \sigma_m^2)^0, (Z(x_0))^0$

2   Sample $(\phi_z, \phi_e)^i$ from $f(\phi_z, \phi_e | y_n, (Z(x_0))^{i-1}, (\sigma_z^2, \sigma_e^2, \sigma_m^2)^{i-1})$

3   Sample $(\sigma_z^2, \sigma_e^2, \sigma_m^2)^i$ from $f(\sigma_z^2, \sigma_e^2 | y_n, (Z(x_0))^{i-1}, (\phi_z, \phi_e)^i)$

4   Sample $Z(x_0)^i$ from $f(Z(x_0) | y_n, (\sigma_z^2, \sigma_e^2, \sigma_m^2)^i, (\phi_z, \phi_e)^i)$

5   Repeat steps 2 to 4 until the chain has converged

6   Approximate $\mu_{|n}(x_0)$ by $\hat{\mu_{|n}} = \frac{\sum_{i=k-m}^{k} Z(x_0)^i}{(k-m)}$ and $\sigma_s | n^2(x_0)$ by $\hat{\sigma_s | n^2} = \frac{\sum_{i=k-m}^{k}(Z(x_0)^i - \hat{\mu_{|n}})^2}{(k-m)-1}$.

The prediction interval for the predictor $\hat{\mu_{|n}}$ can be obtained from the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the $Z(x_0)^i$ samples.

For the assessing of the convergence in step 5, related discussions can be found in (Gilks et al. 1996) and (Gelman, Carlin, Stern, and Rubin 2003). The running time of the algorithm can also be reduced by selecting appropriate starting points. Any prior information can be used to guide the selection of the starting points to increase the algorithm's efficiency. For example, the conditional distribution can be roughly plotted in certain cases to provide some useful insights into the starting point.

To illustrate this general Bayesian metamodeling approach with the MCMC, a numerical example is conducted in the next section. In the Bayesian analysis of these examples, we use general "vague" hyperparameter settings for the priors similar to those adopted in Qian and Wu (2008). Specifically, we chose a "vague" prior of $IG(2,1)$ for $\sigma_z^2, \sigma_e^2, \sigma_m^2$, and the "location-flat" prior $N(0, \sigma_z^2)$ for the β, and $G(2,1)$ for the ϕ_z, ϕ_e.

3   NUMERICAL EXAMPLE

The M/M/1 queue system is a typical stochastic system which is widely studied in the literature. Since the expected waiting time for the M/M/1 system has a closed form, it is easy to compare the model's performance with the true performance. In this stochastic simulation system, we treat traffic rate as the input and the expected waiting time as the output.

Furthermore, in order to study the effect of the noise level on the parameter estimates and predictions, we compare our predictive model with that of the Modified Nugget Effect Kriging (MNEK) model ((Yin, Ng, and Ng 2008)). Since the mathematical forms of the MNEK and the stochastic kriging model are equivalent, we will make the comparison only with MNEK.

In this example, we fix the system arrival rate at 1 and vary the traffic rate. The mean of the expected waiting time can be expressed as a function of traffic rate x:

$$\mu_t = \frac{x}{1-x}$$

For the experimental design, 17 observation points are evenly located at $x = 0.05, 0.10, \ldots, 0.80, 0.85$, where the traffic rate x is considered to be the input of the simulation model. For each observation point, we make 100 replications and take the sample mean and sample variance as the simulation outputs. Based on the 17 sets of input-outputs combinations, we build the MNEK and the Bayesian approach for kriging metamodeling (BK). After that, both models are tested over 1000 evenly distributed prediction points in the sample space. For the modified nugget effect kriging model, the predictor's output is used as the predictive mean and the Mean Squared Prediction Error (MSPE) is used as the predictive variance. For the Bayesian approach for kriging, the predictive mean and predictive variance are approximated over all the MCMC samples.

As the variability at low traffic rate is small, the additional variability in the parameter estimates is small and insignificant. For this reason, we focus the plot in Figure (2) at the higher traffic regions with $x > 0.55$. From Figure (2), we see that the predictive mean given by both MNEK and BK are close to the true mean, while BK gives wider prediction interval than the MNEK. This result can be expected as BK estimates the noise by "borrowing" information from the nearby observed variability
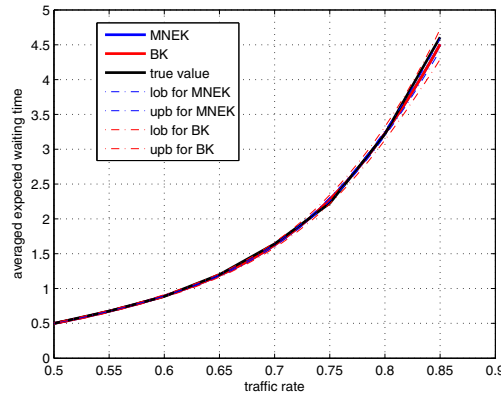
Figure 2: Predictive mean and predictive interval given by MNEK and BK for the M/M/1 example.

and at the same time, takes into account all the parameter estimation uncertainties. This is also aligned with the results provided in (Santner, Williams, and Notz 2003). Furthermore, we can expect that the BK will have a better coverage than the MNEK for the actual simulation runs.
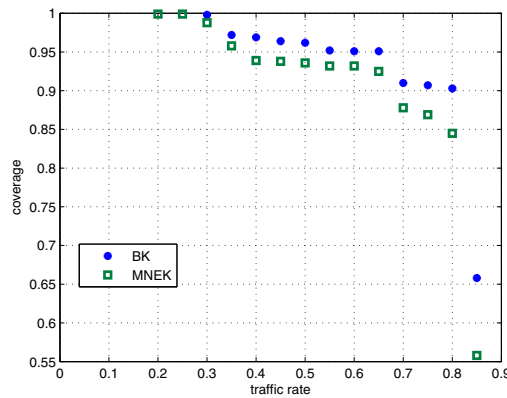

Figure 3: Coverage of the 90% predictive interval for both of the MNEK and BK model for the M/M/1 example.

Figure (3) shows the coverage of the 90% prediction interval for both MNEK and BK models given 1000 actual simulation runs each at the 17 test sites. We can see that the BK has better coverage than the MNEK, especially in the areas with higher variability. Paired t-test are carried out and the results suggest that the difference between the coverage for the MNEK and BK is statistically significant at an alpha level of 0.05. We note that the coverage falls short of the stated level at high traffic level as also observe in Kleijnen and Beers (2010). As for the predictive variance, we see from Fig. (4) that both models underestimate the empirical variance around the area with high variability like from 0.8 to 0.85. BK overestimates the predictive variance in the low traffic region but gives closer results than the MNEK in the high traffic region.

Table 1: Squared difference of the predictive variance given by BK and MNEK for the M/M/1 example.

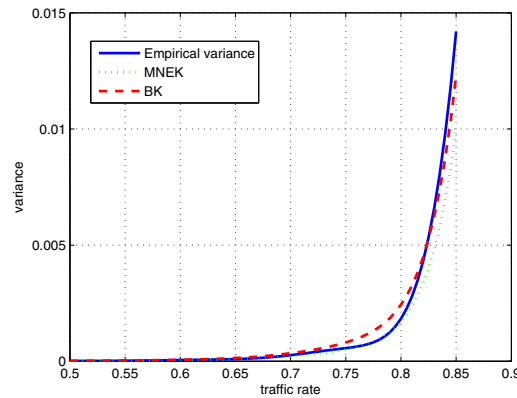| Traffic Rate	MNEK	BK
0.82	8.8E-7	2.9E-8
0.83	3.0E-6	1.1E-7
0.84	8.1E-6	1.1E-6
0.85	1.8E-5	3.7E-6

1061

Figure 4: Predictive variance given by MNEK and BK for the M/M/1 example.

The squared differences between the two models' predictive variance and the empirical variance in the high traffic region are also given in Table (1). We can see the squared difference of the models increases as the traffic rate increases, however BK's squared difference increases at a lower rate compared to MNEK.

4 Conclusion

In this paper, we propose a Bayesian kriging metamodeling approach for stochastic simulation. We model the stochastic simulation outputs as the sum of two different Gaussian random processes with unknown parameters, one to model the mean response surface, and the other the variability. The proposed approach can account for all parameter uncertainties as well as the inherent uncertainty of the simulation model itself. A simple two point example is used to illustrate analytically the basic ideas of the approach. When the ratio of the process variances τ is known, we further derive the predictive distribution. We subsequently propose a MCMC approach to address the prediction for the general case. Using the Bayesian predictive approach, the prediction can explicitly incorporate uncertainties in the model parameters and overcome the problems identified in (Yin, Ng, and Ng 2009).

A comparison between the modified nugget effect model and the proposed Bayesian approach for kriging metamodeling are given based on the M/M/1 simulation system. The numerical results suggest that the predictive variance given by the Bayesian approach is closer to the empirical value than the results provided by modified nugget effect model as the Bayesian predictive approach can account for the parameter uncertainties in the model. However, at high traffic rates, the Bayesian approach still underestimates the variability of the actual simulation model. Follow up work includes applying this predictive modeling approach for designing and planning simulation experiments to more efficiently develop and apply metamodels.

A Proof for Proposition 1

If τ is known, the posterior distribution in (6) can be given as:

$$
\begin{aligned}
f(Z(x_0|Y_n)) &= \iint f(Z(x_0)|Y_n,\beta,\sigma_z^2)f(Y_n|\beta,\sigma_z^2)f(\beta|\sigma_z^2)f(\sigma_z^2)d\beta d\sigma_z^2 \\
&\propto \iint (\sigma_z^2)^{-1/2}\exp(-\frac{1}{2}\frac{Z(x_0)-F^T(x)\beta-r_z^T(x_0)\Sigma^{-1}(Y_n-F(x)\beta)}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)})^2 \\
&\quad (\sigma_z^2)^{-(p+1)/2}\exp(-\frac{1}{2}(Y_n-F^T(x)\beta)^T\Sigma^{-1}(Y_n-F^T(x)\beta)) \\
&\quad (\sigma_z^2)^{-\alpha-n/2-1}\exp(-\frac{1}{2}*(\beta-b_0)^T\frac{V_0^{-1}}{\sigma_z^2}(\beta-b_0))\exp(-\frac{1}{2}\frac{\gamma_z}{\alpha_z})d\beta d\sigma_z^2 \\
&= \iint \exp(-\frac{1}{2}\beta^TA\beta+B^T\beta)\exp(-\frac{1}{2}\frac{\gamma_z}{\alpha_z})(\sigma_z^2)^{-\alpha_z-n/2-3/2-(p+1)/2} \\
&\quad \exp(-\frac{1}{2}(\frac{p'p}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)}+Y_n^T\Sigma^{-1}Y_n+b_0^T\frac{V_0^{-1}}{\sigma_z^2}b_0)+\gamma_z)d\beta d\sigma_z^2 \\
&= \int (\sigma_z^2)^{-\alpha_z-n/2-3/2}\sqrt{\det 2\pi A^{-1}}\exp(\frac{1}{2}B^TA^{-1}B-\frac{1}{2}(\frac{p'p}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)}+Y_n^T\Sigma^{-1}Y_n+b_0^T\frac{V_0^{-1}}{\sigma_z^2}b_0+\gamma_z))d\sigma_z^2
\end{aligned}
$$

where

$$
\begin{aligned}
A &= (\frac{k^Tk}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)}+F^T(x)\Sigma^{-1}F(x)+\frac{V_0^{-1}}{\sigma_z^2}) \\
B &= (\frac{p^Tk}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)}+Y_n^T(x)\Sigma^{-1}F(x)+\frac{b_0V_0^{-1}}{\sigma_z^2}) \\
p &= Z(x_0)-r_z^T(x_0)\Sigma^{-1}Y_n \\
k &= r_z^T(x_0)\Sigma^{-1}Y_n-F^T(x_0) \\
\Sigma &= \sigma_z^2R_z+\sigma_e^2R_e+\sigma_m^2=\sigma_z^2(R_z+\tau R_e)=\sigma_z^2R \\
r_z(x_0) &= \sigma_z^2c_z(x_0)
\end{aligned}
$$

Collecting all terms involving σ_z^2, we have:

$$
\begin{aligned}
f(Z(x_0)|Y_n) &\propto \int (\sigma_z^2)^{-\alpha_z-n/2-3/2}\exp(-\frac{\Omega}{\sigma_z^2})d\sigma_z^2 \\
&\propto \Omega^{-\alpha_z-n/2-1/2}
\end{aligned}
$$

where

$$
\begin{aligned}
\Omega &= (\frac{1}{2}B^TA^{-1}B-\frac{1}{2}(\frac{p'p}{\sigma_z^2-r_z^T(x_0)\Sigma^{-1}r_z(x_0)}+Y_n^T(R_z+\tau R_e)^{-1}Y_n+b_0^TV_0^{-1}b_0+\gamma_z) \\
&= 1+\frac{1}{2\alpha+n}\frac{(Z(x_0)-\mu_p(x_0))^2}{\sigma_p^2(x_0)} \\
\mu_p(x_0) &= F(x_0)\lambda/M+c_z^T(x_0)R^{-1}(y_n-F(x)\lambda/M) \\
\sigma_p^2(x_0) &= \frac{2\gamma_z+y_n^TR^{-1}y_n+b_0^TV_0^{-1}b_0-\lambda^T\lambda/M}{2\alpha_z+n}(1-c_z^T(x_0)R^{-1}c_z(x_0)+h^Th/M) \\
\lambda &= F^TR^{-1}Y_n+b_0V_0^{-1} \\
M &= F^TR^{-1}F+V_0^{-1} \\
h &= r_z^T(x_0)R^{-1}F-F^T(x_0)
\end{aligned}
$$

B Posterior distribution of the parameters

The posterior distribution of the parameters $(\sigma_z^2,\sigma_e^2,\sigma_m^2)$ can be given as:

$$
\begin{aligned}
f(\sigma_z^2,\sigma_e^2,\sigma_m^2|Y_n) &= \iiint f(\beta,\sigma_z^2,\sigma_e^2,\sigma_m^2,\phi_z,\phi_e|Y_n)d\beta d\phi_z d\phi_e \\
&\propto \iiint f(Y_n|\beta,\sigma_z^2,\sigma_e^2,\sigma_m^2,\phi_z,\phi_e)f(\beta|\sigma_z^2)f(\sigma_z^2)f(\sigma_e^2)f(\sigma_m^2)f(\phi_z)f(\phi_e)d\beta d\phi_z d\phi_e
\end{aligned}
$$

The same for the parameters (ϕ_z, ϕ_e):

$$f(\phi_z, \phi_e|Y_n) = \iiint f(\beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e|Y_n) d\beta d\sigma_z^2 d\sigma_e^2 d\sigma_m^2 s$$

$$\propto \iiint f(Y_n|\beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) f(\beta|\sigma_z^2) f(\sigma_z^2) f(\sigma_e^2) f(\sigma_m^2) f(\phi_z) f(\phi_e) d\beta d\sigma_z^2 d\sigma_e^2 d\sigma_m^2$$

The integrand above can be derived given the prior distribution in (8):

$$I(Y_n, \beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) = f(Y_n|\beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) f(\beta|\sigma_z^2) f(\sigma_z^2) f(\sigma_e^2) f(\sigma_m^2) f(\phi_z) f(\phi_e)$$

$$f(Y_n|\beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) \propto (\det \Sigma)^{-1/2} \exp(-\frac{1}{2}(Y_n - F\beta)^T \Sigma^{-1}(Y_n - F\beta))$$

$$f(\beta|\sigma_z^2) \propto (\sigma_z^2)^{-1/2} \exp(-\frac{1}{2}(\beta - b_0)^T \frac{V_0^{-1}}{\sigma_z^2}(\beta - b_0))$$

$$f(\sigma_z^2) \propto \exp(-\frac{\gamma_z}{\sigma_z^2})(\sigma_z^2)^{-\alpha_z+1}$$

$$f(\sigma_e^2) \propto \exp(-\frac{\gamma_e}{\sigma_e^2})(\sigma_e^2)^{-\alpha_e+1}$$

$$f(\sigma_m^2) \propto \exp(-\frac{\gamma_m}{\sigma_m^2})(\sigma_m^2)^{-\alpha_m+1}$$

$$f(\phi_z) \propto \exp(-b_z\phi_z)(\phi_z)^{a_z-1}$$

$$f(\phi_e) \propto \exp(-b_e\phi_e)(\phi_e)^{a_e-1}$$

Rearrange all the parameters,

$$f(\sigma_z^2, \sigma_e^2, \sigma_m^2|Y_n) = \iiint I(Y_n, \beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) d\beta d\phi_z d\phi_e$$

$$\propto \iint (\det M\Sigma)^{-1/2} \exp(\frac{1}{2}\sigma_z^2 \lambda^T M^{-1} \lambda - Y_n^T \Sigma^{-1} Y_n - b_0^T V_0^{-1} b_0 - \frac{\gamma_z}{\sigma_z^2} - \frac{\gamma_e}{\sigma_e^2} - \frac{\gamma_m}{\sigma_m^2})$$

$$(\sigma_z^2)^{1/2-\alpha_z}(\sigma_e^2)^{1-\alpha_e}(\sigma_m^2)^{1-\alpha_m} \exp(-b_z\phi_z - b_e\phi_e)(\phi_z)^{a_z-1}(\phi_z)^{a_e-1} d\phi_z d\phi_e$$

$$f(\phi_z, \phi_e|Y_n) = \iiint I(Y_n, \beta, \sigma_z^2, \sigma_e^2, \sigma_m^2, \phi_z, \phi_e) d\beta d\sigma_z^2 d\sigma_e^2 d\sigma_m^2$$

$$\propto \iint (\det M\Sigma)^{-1/2} \exp(\frac{1}{2}\sigma_z^2 \lambda^T M^{-1} \lambda - Y_n^T \Sigma^{-1} Y_n - b_0^T V_0^{-1} b_0 - \frac{\gamma_z}{\sigma_z^2} - \frac{\gamma_e}{\sigma_e^2} - \frac{\gamma_m}{\sigma_m^2})$$

$$(\sigma_z^2)^{1/2-\alpha_z}(\sigma_e^2)^{1-\alpha_e}(\sigma_m^2)^{1-\alpha_m} \exp(-b_z\phi_z - b_e\phi_e)(\phi_z)^{a_z-1}(\phi_z)^{a_e-1} d\sigma_z^2 d\sigma_e^2 d\sigma_m^2$$

where

$$\lambda = F^T R^{-1} Y_n + b_0 V_0^{-1}, M = F^T R^{-1} F + V_0^{-1}$$

As for the direct sampling from these two posterior distribution can be difficult, we adopt the rejection-acceptance method for the posterior distribution sampling within the Gibbs loop for the MCMC implement.

C Details of the two-point example

Assuming the non informative prior (Jefferys Prior):

$$p(\beta) = 1, p(\sigma_z^2) = 1/\sigma_z^2, p(\tau) = 1/(2(1+\tau)^2)$$

With the covariance matrix Σ, correlation vector $r(x_0)$, observation vector Y_n, and regressors $F(x)$ and $F(x_0)$ given in Eq. (7), following the steps in Proposition 1, we get:

$$
\begin{aligned}
p &= Z(x_0) - r_z^T(x_0)\Sigma^{-1}Y_n = \frac{y_1(1+\rho) + y_2(1+\rho) - 2Z(x_0)(1+\rho+\tau)}{2(1+\rho+\tau)} \\
k &= r_z^T(x_0)\Sigma^{-1}Y_n - F^T(x_0) = -\frac{\tau}{1+\rho+\tau} \\
A &= \left(\frac{k^Tk}{\sigma_z^2 - r_z^T(x_0)\Sigma^{-1}r_z(x_0)} + F^T(x)\Sigma^{-1}F(x) + \frac{V_0^{-1}}{\sigma_z^2}\right) = -\frac{1}{\sigma_z^2}\frac{2(-1+\rho-\tau)}{-1+\rho^2-2\tau} \\
B &= \left(\frac{p^Tk}{\sigma_z^2 - r_z^T(x_0)\Sigma^{-1}r_z(x_0)} + Y_n^T(x)\Sigma^{-1}F(x) + \frac{b_0V_0^{-1}}{\sigma_z^2}\right) = -\frac{1}{\sigma_z^2}\frac{2(y_1+y_2-y_1\rho-y_2\rho+2Z(x_0\tau))}{1-\rho^2+2\tau}
\end{aligned}
$$

Collecting all the terms with τ, the posterior distribution can be simplified to:

$$
\begin{aligned}
f(Z(x_0)|Y_n) &\propto \int \sqrt{2\pi A^{-1}}\left(\frac{1}{2}B^TA^{-1}B - \frac{1}{2}\left(\frac{p^Tp}{1-c_z^TR^{-1}c_z} + Y_n^TR^{-1}Y_n\right)\right)^{-1/2}d\tau \\
&\propto \sqrt{\frac{1-\rho^2}{y_1^2 + y_2^2 - 2(y_1+y_2)Z(x_0) + 2Z(x_0)^2}}
\end{aligned}
$$

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. Stochastic kriging for simulation metamodeling. *Operations Research (to appear)*.

Cressie, N. 1993. *Statistics for spatial data*. revised ed. New York: Wiley.

den Hertog, J. P. C. Kleijnen, and A. Y. D. Siem. 2005. The correct kriging variance estimated by bootstrapping. *Journal of the Operational Research Society* 165(3):826–834.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian data analysis*. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov chain monte carlo in practice*. 1st ed. London: Chapman & Hall.

Gupta, A., D. Yu, L. Xu, and T. Reinikainen. 2006. Optimal parameter selection for electronic packaging using sequential computer simulations. *Journal of Manufacturing Science and Engineering, Transactions of the ASME* 128:705–715.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng. 2006. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization* 34(3):441–466.

Jeffreys, H. 1961. *Theory of probability*. London: Oxford University Press.

Kleijnen, J. P. C., and W. C. M. V. Beers. 2005. Robustness of kriging when interpolating in random simulation with heterogeneous variances: Some experiments. *European Journal of Operational Research* 30:102–119.

Kleijnen, J. P. C., and W. C. M. V. Beers. 2010. Monotonicity-preserving bootstrapped kriging metamodels for expensive simulations. *Working paper*.

Li, Y. F., S. H. Ng, M. Xie, and T. N. Goh. 2010. A systematic comparison of metamodeling techniques for simulation optimization in decision support systems. *Applied Soft Computing (to appear)*.

Matheron, G. 1963. Principle of geostatistics. *Economic Geology* 58:1246–1266.

Pham, T., and M. Wagner. 1994. Filtering noisy images using kriging, signal processing and its applications. In *Proceedings of the Fifth International Symposium*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 184–191. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Qian, P. Z. G., and C. F. J. Wu. 2008. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics* 50(2):192–204.

Roshan, J. V. 2006. Limit kriging. *Technometrics* 48 (4): 458–466.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The design and analysis of computer experiments*. New York: Springer.

Simpson, T., J. Peplinski, P. Koch, and J. Allen. 2001. Metamodels for computer-based engineering design: Survey and recommendations. *Engineering with Computers* 17:129–150.

Wang, S., W. Chen, and K. L. Tsui. 2009. Bayesian validation of computer models. *Technometrics* 51 (4): 439–451.

Wu, H., and F. Sun. 2007. Adaptive kriging control of discrete-time nonlinear systems. *Control Theory & Applications* 1 (3): 646–656.

Yin, J., S. H. Ng, and K. M. Ng. 2008. Kriging model with modified nugget effect for random simulation with heterogeneous variances. In *In Proceedings of IEEE International Conference on Industrial Engineering and*

Engineering Management, ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 674–685. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yin, J., S. H. Ng, and K. M. Ng. 2009. A study on the effect of parameter estimation on the kriging model's prediction error in stochastic simulation. In *In Proceedings of the 2009 Winter Simulation Conference*, ed. M. Rossetti, R. Hill, B. Johnson, A. Dunkin, and R. Ingalls, 674–685. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zimmerman, D. L., and N. Cressie. 1992. Mean squared error prediction error in the spatial linear model with estimated co-variance parameters. *Annals of the institute of statistical mathematics* 44 (1): 27–43.

AUTHOR BIOGRAPHIES

JUN YIN is a Ph.D candidate in the Department of Industrial and Systems Engineering at the National University of Singapore. He received his B. Eng. in Automation from University of Science and Technology of China in 2005. His research in terests include computer simulation, design of experiments and metamodeling. His email address is <yin-jun@nus.edu.sg>.

SZU HUI NG is an Assistant Professor in the Department of Industrial and Systems Engineering at the National University of Singapore. She holds B.S., M.S. and Ph.D. degrees in Industrial and Operations Engineering from the University of Michigan. Her research interests include quality and reliability modeling and analysis, design of experiments and simulation analysis. Her e-mail and web addresses are <isensh@nus.edu.sg> and <http://www.ise.nus.edu.sg/staff/ngsh/index.html>.

KIEN MING NG is an Assistant Professor in the Department of Industrial and Systems Engineering at the National University of Singapore. He obtained his PhD in Management Science and Engineering from Stanford University in 2002. His research interests are in optimization and numerical algorithms, as well as operations research applications in logistics. He can be contacted by email at <isenkm@nus.edu.sg>.