Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

STOCHASTIC KRIGING WITH QUALITATIVE FACTORS

Xi Chen

Statistical Sciences and Operations Research
Virginia Commonwealth University
Richmond, VA 23284, USA

Kai Wang and Feng Yang

Industrial and Management Systems Engineering
West Virginia University
Morgantown, WV 26506, USA

ABSTRACT

Stochastic kriging (SK) has been studied as an effective metamodeling technique for approximating the mean response surface implied by a stochastic simulation. Until recently, it has only been applied to simulation experiments with continuous decision variables or factors. In this paper, we propose a new method called stochastic kriging with qualitative factors (SKQ) that extends stochastic kriging to a broader scope of applicability. SKQ is able to build metamodels for stochastic simulations that have both quantitative (continuous) and qualitative (categorical) factors. To make this extension, we introduce basic steps of constructing valid spatial correlation functions for handling correlations across levels of qualitative factors. Two examples are used to demonstrate the advantages of SKQ in aggregating information from related response surfaces and metamodeling them simultaneously, in addition to maintaining SK's ability of effectively tackling the impact of simulation errors.

1 INTRODUCTION

In applications of operations research and management sciences, simulation models with both qualitative and quantitative decision variables or factors are frequently used to solve real-world problems. For instance, in production planning of semiconductor manufacturing systems, simulation is commonly employed to determine the appropriate release rate of raw materials (a continuous variable) and dispatch rules (a qualitative variable), in an effort to optimize the system performance, which is typically measured in terms of the system throughput and products' cycle time. For some complicated simulation tasks, it may take days or even weeks to complete a single simulation run, which potentially limits the usefulness of simulation in providing support for real-time decision making. To mitigate this deficiency, carefully designed simulation experiments can be employed to build metamodels to approximate some aspects of system performance without running extensive simulations. Stochastic kriging is one of the tools proposed recently for representing stochastic simulation response surfaces. Despite its success achieved so far, stochastic kriging assumes that all the decision variables or inputs to simulation experiments are continuous. Therefore, an important yet underdeveloped topic is how to construct stochastic kriging metamodels for stochastic simulations with both quantitative and qualitative decision variables. The purpose of this paper is to make the first attempt to extend the standard stochastic kriging metamodels to address the question above.

Methods for building Gaussian process (GP) based metamodels with quantitative and qualitative factors have been developed for deterministic computer experiments. We give a brief summary of the recent literature and refer the interested reader to references therein. The key to any development is to construct valid correlation functions for GP models with both types of factors. Qian et al. (2008) established a general approach for constructing unrestrictive and restrictive correlation functions and also developed a corresponding iterative procedure for model parameter estimation. Zhou et al. (2011) proposed to use the unrestrictive correlation structure with the hypersphere decomposition which helps simplify the parameter estimation procedure used by Qian et al. (2008). Han et al. (2009) introduced a hierarchical

Bayesian model with conditional Gaussian stochastic process components for computer experiments having quantitative and qualitative inputs. It seems natural and straightforward for us to take advantage of the aforementioned advances and apply them to generalize the stochastic kriging metamodeling framework. However, some unique features enjoyed by a stochastic simulation may disguise important problems worth further investigation. One of such problems absent in deterministic computer experiments is the impact of stochastic simulation errors on the performance of resulting metamodels. In this paper we take the first few steps in seeking and further investigating interesting questions arisen in the process of extending the stochastic kriging framework.

The remainder of the paper is organized as follows. Section 2 outlines the standard stochastic kriging framework. Section 3 introduces the main ingredient for building SKQ metamodels—constructing valid correlation functions with both quantitative and qualitative factors based on Qian et al. (2008) and Zhou et al. (2011). Section 4 presents two examples to demonstrate the competitive performance of the SKQ metamodels subject to varying levels of stochastic simulation errors, and their advantage of incorporating output information from relevant response surfaces that have "similar" functional behavior to do prediction. Section 5 concludes the paper.

2 STANDARD STOCHASTIC KRIGING

This section gives a brief review on stochastic kriging which was first introduced by Ankenman et al. (2010). In stochastic kriging the simulation's output on the jth replication at design point $\mathbf{x}$ is represented as

$$\mathscr{Y}_j(\mathbf{x}) = \mathsf{Y}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \beta + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) \tag{1}$$

In (1) $\mathsf{Y}(\mathbf{x})$ represents the underlying true response surface at design point $\mathbf{x}$, a vector of input or decision variables in $\mathbb{R}^d$. The $p \times 1$ vector of functions $\mathbf{f}(\mathbf{x})$ is typically assumed to be known, whereas the corresponding vector of coefficients β often needs to be estimated. The constant mean model $\mathbf{f}(\mathbf{x})^\top \beta = \beta_0$ has been widely adopted in the kriging literature since it has been reported performing sufficiently well for most applications in practice. The term M represents a realization of a mean zero stationary *Gaussian random field*; that is, we can think of M as being randomly sampled from a space of functions mapping $\mathbb{R}^d \to \mathbb{R}$, and the functions in that space exhibit spatial correlations. Finally, $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ are the independent and identically distributed mean zero simulation errors incurred on each replication at design point $\mathbf{x}$. As in Ankenman et al. (2010), we sometimes refer to $\mathsf{M}(\mathbf{x})$ and $\varepsilon_j(\mathbf{x})$ as the extrinsic and intrinsic uncertainties at design point $\mathbf{x}$.

An experiment for building a stochastic kriging metamodel consists of running n_i simulation replications at $\mathbf{x}_i, i = 1, 2, \ldots, k$, assuming that there are in total k design points. Notice that n_i's can differ at distinct design points. Following (1), we can express the sample average simulation output at $\mathbf{x}_i$ across replications as

$$\bar{\mathscr{Y}}(\mathbf{x}_i) \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} \mathscr{Y}_j(\mathbf{x}_i) \;=\; \mathbf{f}(\mathbf{x}_i)^\top \beta + \mathsf{M}(\mathbf{x}_i) + \frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i);$$

and use $\bar{\mathscr{Y}} = \left(\bar{\mathscr{Y}}(\mathbf{x}_1), \bar{\mathscr{Y}}(\mathbf{x}_2), \ldots, \bar{\mathscr{Y}}(\mathbf{x}_k)\right)^\top$ to denote the $k \times 1$ vector of averaged simulation outputs at all k design points; correspondingly, we can represent the vector of averaged simulation errors by $\bar{\varepsilon} = (\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_k))^\top$, where $\bar{\varepsilon}(\mathbf{x}_i) = n_i^{-1}\sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i), i = 1, 2, \ldots, k$.

2.1 The Extrinsic and Intrinsic Variance Structures

Define $\Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[\mathsf{M}(\mathbf{x}), \mathsf{M}(\mathbf{x}')] = \mathrm{Cov}[\mathsf{Y}(\mathbf{x}), \mathsf{Y}(\mathbf{x}')]$ to be the spatial covariance of points $\mathbf{x}$ and $\mathbf{x}'$ implied by the extrinsic spatial correlation model; and let the $k \times k$ matrix Σ_{M} represent the extrinsic spatial variance-covariance matrix of the k design points such that the ijth entry of Σ_{M} gives the spatial covariance of the ith and jth design points, $i, j = 1, 2, \ldots, k$. Finally, let $\mathbf{x}_0$ be the prediction point, and define $\Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)$

to be the $k \times 1$ vector that contains the extrinsic spatial covariances between $\mathbf{x}_0$ and each of the k design points; that is, $\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot) = (\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0),\mathsf{M}(\mathbf{x}_1)],\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0),\mathsf{M}(\mathbf{x}_2)],\ldots,\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0),\mathsf{M}(\mathbf{x}_k)])^\top$. Given that M is stationary, Σ_{M} and $\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)$ are assumed to take the following forms:

$$\Sigma_{\mathsf{M}} = \tau^2\mathbf{R} = \tau^2\begin{pmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \ldots & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2\mathbf{r}_0 = \tau^2\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{pmatrix} \tag{2}$$

where $\tau^2 > 0$ denotes the extrinsic spatial variance. The $k \times k$ matrix $\mathbf{R}$ contains the pairwise spatial correlations of the k design points. The $k \times 1$ vector $\mathbf{r}_0$ gives respective spatial correlations between the k design points and the prediction point $\mathbf{x}_0$. To do metamodel estimation and prediction we need to impose structure on the form of the spatial correlation function. There are several choices of the spatial correlation functions available, see, for instance, Santner et al. (2003) and Qian et al. (2008). One particular choice that is quite popular in practice is the exponential correlation function

$$K(\mathbf{x}_i,\mathbf{x}_h) = \exp\left\{\sum_{\ell=1}^{d} -\theta_\ell|x_{i\ell} - x_{h\ell}|^p\right\}, \tag{3}$$

where $\theta_\ell \geq 0$ controls the roughness of the response surface and θ_ℓ, $\ell = 1,2,\ldots,d$ for different coordinates are not necessarily identical. The parameter p is in $(0,2]$. In particular, the sample paths of a spatial process M are infinitely differentiable if the popular Gaussian correlation function (correspondingly, $p = 2$) is used. In this case, the spatial correlation between design point $\mathbf{x}_i$ and prediction point $\mathbf{x}_0$ becomes $r_i = K(\mathbf{x}_i,\mathbf{x}_0)$, and the spatial correlation between two design points $\mathbf{x}_h$ and $\mathbf{x}_i$ follows as $r_{ih} = K(\mathbf{x}_i,\mathbf{x}_h)$ with $p = 2$.

What distinguishes stochastic kriging from kriging for deterministic computer experiments (e.g., Sacks et al. 1989; Santner et al. 2003) is that the former accounts for the sampling variability inherent in a stochastic simulation. Let Σ_ε be the $k \times k$ intrinsic variance-covariance matrix for $\bar{\varepsilon}$. If common random numbers are not used in driving a simulation experiment, then Σ_ε takes a diagonal matrix, which can be specified as

$$\Sigma_\varepsilon = \mathrm{diag}\{\sigma_1^2/n_1, \sigma_2^2/n_2, \ldots, \sigma_k^2/n_k\},$$

where $\sigma_i^2 = \mathrm{Var}[\varepsilon_j(\mathbf{x}_i)]$ is the intrinsic variance at design point $\mathbf{x}_i, i = 1,2,\ldots,k$.

2.2 Prediction by Stochastic Kriging

Chen et al. (2012) show that when β is unknown but Σ_ε and the spatial parameters are known, the MSE-optimal predictor provided by stochastic kriging takes the following form

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top\widehat{\beta} + \Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\Sigma^{-1}(\bar{\mathscr{Y}} - \mathbf{F}\widehat{\beta}) \ ,$$

where the rows of $k \times p$ matrix $\mathbf{F}$ are, respectively, $\mathbf{f}(\mathbf{x}_1)^\top, \mathbf{f}(\mathbf{x}_2)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top$, $\Sigma = \Sigma_{\mathsf{M}} + \Sigma_\varepsilon$ and $\widehat{\beta} = \left(\mathbf{F}^\top\Sigma^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\Sigma^{-1}\bar{\mathscr{Y}}$ is the generalized least squares estimator of β. The corresponding mean squared error (MSE) of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$ follows as

$$\mathrm{MSE}(\widehat{\mathsf{Y}}(\mathbf{x}_0)) = \Sigma_{\mathsf{M}}(\mathbf{x}_0,\mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\Sigma^{-1}\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot) + \eta^\top(\mathbf{F}^\top\Sigma^{-1}\mathbf{F})^{-1}\eta \ , \tag{4}$$

where $\eta = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top\Sigma^{-1}\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)$. For an actual implementation of stochastic kriging for prediction, one needs to estimate the necessary spatial parameters as well as Σ_ε. The standard approach is first to estimate the intrinsic variance-covariance matrix Σ_ε using the sample covariance matrix $\widehat{\Sigma}_\varepsilon$ whose diagonal entries are

$$\widehat{\Sigma}_{\varepsilon_{i,i}} := \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\left(\mathscr{Y}_j(\mathbf{x}_i) - \bar{\mathscr{Y}}(\mathbf{x}_i)\right)^2 \ , \ i = 1,2,\ldots,k \ . \tag{5}$$

Then by substituting $\widehat{\Sigma}_\varepsilon$ for Σ_ε, we can estimate the spatial parameters (θ and τ^2, if the Gaussian spatial correlation function is used) and β through maximizing the resulting log-likelihood function.

3 STOCHASTIC KRIGING WITH QUALITATIVE FACTORS

In this section, we extend the standard stochastic kriging to incorporate qualitative factors into the design space. We adopt the notation established in Qian et al. (2008) and Zhou et al. (2011) whenever possible. Recall (1), in which the vector of decision variables $\mathbf{x}$ is assumed to be in $\mathbb{R}^d$. Now let us denote the vector of decision variables by $\mathbf{w} = (\mathbf{x}^\top, \mathbf{z}^\top)^\top$, in which $\mathbf{x} \in \mathbb{R}^d$ represents the continuous factors and $\mathbf{z} = (z_1, z_2, \ldots, z_J)^\top$ are qualitative factors with z_j having m_j levels, $j = 1, 2, \ldots, J$. Similar to (1), the simulation's output at design point $\mathbf{w}$ on the ℓth simulation replication can be modeled as

$$\mathscr{Y}_\ell(\mathbf{w}) = \mathsf{Y}(\mathbf{w}) + \varepsilon_\ell(\mathbf{w}) = \mathbf{f}(\mathbf{w})^\top \beta + \mathsf{M}(\mathbf{w}) + \varepsilon_\ell(\mathbf{w}), \tag{6}$$

where as before $\mathbf{f}(\mathbf{w})$ is a $p \times 1$ vector of known functions and β is a vector of unknown parameters of appropriate dimension. As for M, we assume that the values corresponding to different levels of a qualitative factor are drawn from Gaussian random processes with "similar" spatial correlation structures and magnitudes of variation. Everything discussed in Section 2.1 regarding stochastic simulation errors applies to the $\varepsilon_\ell(\mathbf{w})$'s. Furthermore, the intrinsic variance-covariance matrix Σ_ε can still be estimated by the sample covariance matrix $\widehat{\Sigma}_\varepsilon$ given in (5). The key to extend stochastic kriging to incorporate qualitative factors lies in constructing valid spatial correlation functions for M, on which we are to elaborate next.

3.1 Spatial Correlation Functions for Stochastic Kriging with Qualitative Factors

Consider the general case with J qualitative factors $\mathbf{z} = (z_1, z_2, \ldots, z_J)^\top$ with z_j having m_j levels, $j = 1, 2, \ldots, J$. Following the discussion in Qian et al. (2008), we propose to use a spatial correlation function for M constructed as follows. For a pair of design points $\mathbf{w}_i$ and $\mathbf{w}_h$, $i, h = 1, 2, \ldots, k$,

$$\mathrm{Corr}[\mathsf{M}(\mathbf{w}_i), \mathsf{M}(\mathbf{w}_h)] = \left[\prod_{j=1}^{J} \tau_{j, z_{ij}, z_{hj}}\right] \cdot K(\mathbf{x}_i, \mathbf{x}_h) \ . \tag{7}$$

In particular, we observe that the correlation function above consists of two parts, respectively, dedicated to modeling the spatial correlations regarding the qualitative and quantitative (continuous) decision variables. More specifically, the second term on the right-hand side of (7) is just $K(\mathbf{x}_i, \mathbf{x}_h)$ as given in (3); while the first term is the new ingredient added for handling spatial correlations across different levels of qualitative factors. As noted in Qian et al. (2008), the following interpretations can help us develop intuition about the proposed correlation function: The parameter $\tau_{j, z_{ij}, z_{hj}}$ measures the similarity at any two design points $\mathbf{w}_i$ and $\mathbf{w}_h$ that differ only on the values of qualitative factors z_j; the correlation parameters $\theta_\ell, \ell = 1, 2, \ldots, d$ in $K(\cdot, \cdot)$, on the other hand, assess the roughness of the sample path of the Gaussian process M given the same values of $\mathbf{z}$ for the qualitative factors.

Several correlation functions for the qualitative factors have been proposed in the literature. We briefly introduce some of them below and refer the interested reader to Qian et al. (2008) and Zhou et al. (2011) for full details.

- Isotropic (or exchangeable) correlation functions (EC): For qualitative factor $j = 1, 2, \ldots, J$, with $\phi_j > 0$,

  $$\tau_{j, z_{ij}, z_{hj}} = \exp\{-\phi_j I[z_{ij} \neq z_{hj}]\} \ ,$$

  where $I[A]$ is an indicator function that takes 1 if event A is true and 0 otherwise. The isotropic correlation function assumes that the m_j levels of z_j are of isotropic nature; that is, $\tau_{j, z_{ij}, z_{hj}} = c_j$ for all $z_{ij} \neq z_{hj}$.

- Multiplicative correlation functions (MC): For qualitative factor $j = 1, 2, \ldots, J$, the following $\tau_{j,z_{ij},z_{hj}}$ allows different pairs of z_j levels to have different correlations

$$\tau_{j,z_{ij},z_{hj}} = \exp\big\{-(\phi_{ij}+\phi_{hj})I[z_{ij} \neq z_{hj}]\big\} \ .$$

- Group correlation functions and correlation functions for ordinal qualitative factors. The latter correlation function is particularly suitable for dealing with ordinal qualitative factors; The interested reader is referred to Section 4.4 of Qian et al. (2008) for details. We note that in this paper attention is restricted to qualitative factors that are nominal but not ordinal.

Zhou et al. (2011) propose to use a new type of correlation functions for the qualitative factors which are nominal but not ordinal, namely, the unrestrictive correlation functions (UC). UC functions are constructed through the hypersphere decomposition. Recall that a $(d+J)\times 1$ vector of decision variables $\mathbf{w} = (\mathbf{x}^\top, \mathbf{z}^\top)^\top$ has $\mathbf{z} = (z_1, z_2, \ldots, z_J)^\top$ consisting of all the qualitative factors, with z_j having m_j levels, $j = 1, 2, \ldots, J$.

- The general format of UC. Let $m = \prod_{j=1}^{J} m_j$ and let $c_1, c_2, \ldots, c_m$ denote the m categories, corresponding to all level combinations of the qualitative factors in $\mathbf{z}$. Hence, we can rewrite $\mathbf{w}$ as a $(d+1)\times 1$ vector $\mathbf{w} = (\mathbf{x}^\top, c_q)^\top, q = 1, 2, \ldots, m$ to denote all the factors involved. Then following (7) we have the spatial correlation between two design points $\mathbf{w}_i$ and $\mathbf{w}_h$ can be given as

$$\mathrm{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_h)] = \tau_{c_i,c_h} K(\mathbf{x}_i, \mathbf{x}_h) \ ,$$

where $\tau_{c_i,c_h} = \tau_{c_h,c_i}$ gives the cross-correlation between categories c_i and c_h. If we adopt the Gaussian correlation function for the quantitative factors as we will do in the forthcoming numerical examples in Section 4, the above correlation equation turns into

$$\mathrm{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_h)] = \tau_{c_i,c_h} \exp\left\{\sum_{\ell=1}^{d} -\theta_\ell |x_{i\ell} - x_{h\ell}|^2\right\} \ . \tag{8}$$

We are ready to construct a correlation matrix $\mathbf{T} = [\tau_{r,s}]_{m\times m}$ by the hypersphere decomposition through two steps as prescribed in Zhou et al. (2011). Notice that the following construction guarantees us a positive definite matrix $\mathbf{T}$ with unit diagonal elements.

Step 1. Apply a Cholesky decomposition to the matrix $\mathbf{T}$ and obtain the lower triangular matrix with strictly positive diagonal entries $\mathbf{L}$ such that $\mathbf{T} = \mathbf{L}\mathbf{L}^\top$.

Step 2. Each row vector $(l_{r,1}, l_{r,2}, \ldots, l_{r,r})$ in $\mathbf{L}$ is modeled as the coordinates of a surface point on an r-dimensional unit hypersphere described as follows. For $r = 1$, let $l_{1,1} = 1$ and for $r = 2, \ldots, m$, the row elements are specified through the following spherical coordinate system

$$\begin{cases} l_{r,1} &= \cos(\phi_{r,1}), \\ l_{r,s} &= \sin(\phi_{r,1}) \ldots \sin(\phi_{r,s-1}) \cos(\phi_{r,s}), \ s = 2, \ldots, r-1, \\ l_{r,r} &= \sin(\phi_{r,1}) \ldots \sin(\phi_{r,r-1}) \sin(\phi_{r,r-1}), \end{cases}$$

where $\phi_{r,s}$ is in the parameter set $\Phi = \{\phi_{r,s} \in (0,\pi), s = 1, 2, \ldots, r-1; r = 1, 2, \ldots, m\}$.
One advantage among several of the hypersphere decomposition that distinguishes this method is that given $\phi_{r,s} \in (0,\pi)$, the entries $\tau_{r,s}$ in $\mathbf{T}$ can take both positive and negative values. This feature enables us to flexibly model different correlations across the categories (i.e., all levels of all the qualitative factors). The general format of UC as described above, however, does have a drawback. We observe that the cardinality of the parameter set Φ, $|\Phi|$, equals $m(m-1)/2$ for the general format of UC; and this number can get considerably large if there are multiple qualitative factors in the design space, each of which having a number of levels. This drawback can be alleviated by using the product format of UC as specified below.

- The product format of UC. If multiple qualitative factors are under consideration, i.e., $J > 1$, one can use the following product form of the correlation function in place of the one in (8),

$$\text{Corr}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_h)] = \left[\prod_{j=1}^{J} \tau_{j,z_{ij},z_{hj}}\right] \exp\left\{\sum_{\ell=1}^{d} -\theta_\ell |x_{i\ell} - x_{h\ell}|^2\right\}, \tag{9}$$

  where separate correlation matrices $\mathbf{T}_j, j = 1,2,\ldots,J$ are created for each of the J qualitative factors. The matrix $\mathbf{T}_j = [\tau_{j,r,s}], r,s = 1,2,\ldots,m_j$ is modeled by using the parameterization given for the general format of UC. It is easy to tell that although the product format given in (9)) is not as flexible as the general format given in (8), it can reduce $|\Phi|$ to $\sum_{j=1}^{J} m_j(m_j-1)/2$, which could be a substantial saving if each qualitative factor has multiple levels and J is large.

Given design points $\{\mathbf{w}_i\}_{i=1}^k$, by using the aforementioned approaches, i.e., EC, MC and UC, etc., we are able to construct the spatial covariance matrix $\Sigma_{\mathsf{M}} = \tau^2\mathbf{R}$ and the vector $\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2\mathbf{r}_0$ specified in (2) correspondingly.

3.2 Maximum Likelihood Estimation and Prediction for Stochastic Kriging with Qualitative Factors

We first estimate the intrinsic variance-covariance Σ_ε by (5) and substitute $\widehat{\Sigma}_\varepsilon$ into the likelihood function $\mathscr{L}$ to obtain MLEs for τ^2, θ, Φ and β. Given a choice of τ^2, θ, Φ and a given $\widehat{\Sigma}_\varepsilon$, the value of $\widehat{\beta}(\tau^2,\theta,\Phi)$ that maximizes $\mathscr{L}(\tau^2,\theta,\Phi)$ is given by $\widehat{\beta}(\tau^2,\theta,\Phi) = \left(\mathbf{F}^\top\widehat{\Sigma}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\widehat{\Sigma}^{-1}\bar{\mathscr{Y}}$ with $\widehat{\Sigma} = \Sigma_{\mathsf{M}}(\tau^2,\theta,\Phi) + \widehat{\Sigma}_\varepsilon$; it follows that the MLEs $\widehat{\tau^2}$, $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_d)^\top$, $\widehat{\Phi} = \{\widehat{\phi}\}$ can be obtained by maximizing the log-likelihood function specified below subject to $\theta > 0$ and $\tau^2 > 0$:

$$\ln\mathscr{L}(\tau^2,\theta,\Phi) = -\frac{1}{2}\left(k\ln(2\pi) + \ln(|\widehat{\Sigma}|) + \widetilde{\mathscr{Y}}^\top\widehat{\Sigma}^{-1}\widetilde{\mathscr{Y}}\right) \propto -\frac{1}{2}\left(\ln(|\widehat{\Sigma}|) + \widetilde{\mathscr{Y}}^\top\widehat{\Sigma}^{-1}\widetilde{\mathscr{Y}}\right),$$

where $\widetilde{\mathscr{Y}} = \bar{\mathscr{Y}} - \mathbf{F}\widehat{\beta}(\tau^2,\theta,\Phi)$ and $|\widehat{\Sigma}|$ denotes the determinant of $\widehat{\Sigma}$. In fact, we optimize through minimizing $-\ln\mathscr{L}$ using the Matlab's nonlinear optimization function `fmincon`. Notice that the parameter p ($p > 0$) should also be included in the log-likelihood function if an exponential correlation function is used, other than the Gaussian correlation function, in which case $p = 2$.

Prediction through stochastic kriging with qualitative factors (SKQ) can be achieved readily following the lines developed in Section 2.2. Since Zhou et al. (2011) demonstrate the advantage of UC over EC and MC for modeling deterministic computer experiments through numerical comparisons, for the sake of brevity, we only consider implementing SKQ with UC (the product format) in the numerical examples in the next section.

4 NUMERICAL EXAMPLES

In this section, we demonstrate the potential of stochastic kriging with qualitative factors (SKQ) in exploiting information from correlated response surfaces and its ability of providing adequate approximations to multiple response surfaces simultaneously.

4.1 Example 1: An Example with Both Positive and Negative Cross-Correlations

The example in this section is constructed from the example given in Section 5.1 of Zhou et al. (2011). Consider the following experiment with one quantitative factor x and one qualitative factor z with three levels. At design point $\mathbf{w}$, we have $\mathbf{w} = (x,z)^\top$ with $x \in [0,1]$ and $z \in \{1,2,3\}$ and the true response surface takes the following form:

$$\mathsf{Y}(\mathbf{w}) = \begin{cases} 5\cos(6.8\pi x/2) & \text{if } z = 1, \\ -5\cos(7\pi x/2) & \text{if } z = 2, \\ 5\cos(7.2\pi x/2) & \text{if } z = 3. \end{cases}$$

The three true response surfaces corresponding to $z = 1,2,3$ are shown in Figure 1.

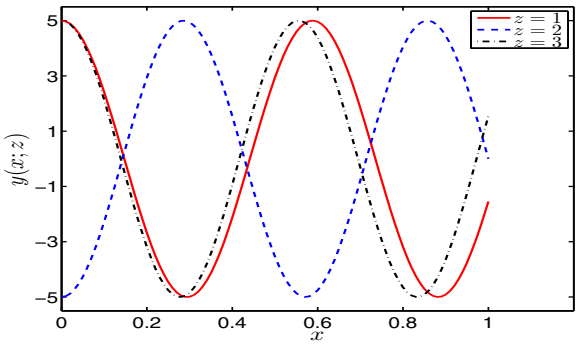

Figure 1: The true response surfaces correspond to $z = 1,2,3$ for example 1.

Experiment design. For each level of z, we use a 10-point design for the continuous factor that is composed of a Latin-hypercube sample (LHS) of 8 points in $[0,1]$ plus the two endpoints 0 and 1; hence there are 30 design points in total. At the ith design point, the simulation output at $\mathbf{w}_i$ on the jth replication is generated according to the following model:

$$\mathscr{Y}_j(\mathbf{w}_i) = \mathsf{Y}(\mathbf{w}_i) + \varepsilon_j(\mathbf{w}_i) \ , \ i = 1,2,\ldots,30,$$

where $\varepsilon_j(\mathbf{w}_i), j = 1,2,\ldots,n$ are i.i.d. $\mathrm{N}(0,\gamma)$, and n denotes the common number of simulation replications applied to all design points. We fix $n = 100$ in this experiment. To see the impact of stochastic simulation errors, we control the magnitude of the intrinsic variance by varying the value of γ in $\{1,25,50,100\}$, which corresponds γ/n in $\{0.01,0.25,0.5,1\}$. We are interested in the performance of SKQ in comparison to two other alternative methods, respectively, the standard stochastic kriging applied to fitting response surfaces for $z = 1,2,3$ separately (SK) and UC proposed by Zhou et al. (2011) for deterministic computer experiments. To this end, we select 50 check-points $\{\mathbf{w}_\ell^z\}_{\ell=1}^{50}$ with their continuous factor equally spaced in $[0,1]$ for each level of z, that is, $\mathbf{w}_\ell^z = (x_\ell, z)^\top$ with $x_\ell = (\ell-1)/49, \ell = 1,2,\ldots,50$ for $z = 1,2,3$, respectively. For evaluation of the predictive performance regarding approximating individual surface that corresponds to $z = i$, we use the estimated root mean squared error (ERMSE_i),

$$\mathrm{ERMSE}_i = \sqrt{\frac{1}{50}\sum_{\ell=1}^{50}\left(\widehat{y}(\mathbf{w}_\ell^i) - \mathsf{Y}(\mathbf{w}_\ell^i)\right)^2} \ ,$$

where $\mathbf{w}_\ell^i = (x_\ell, i)^\top, i = 1,2,3$; the predicted value is denoted by $\widehat{y}(\cdot)$, and $\mathsf{Y}(\cdot)$ stands for the true response. The overall predictive performance across the three surfaces is examined through ERMSE defined as follows:

$$\mathrm{ERMSE} = \sqrt{\frac{1}{150}\sum_{i=1}^{3}\sum_{\ell=1}^{50}\left(\widehat{y}(\mathbf{w}_\ell^i) - \mathsf{Y}(\mathbf{w}_\ell^i)\right)^2} \ .$$

Results. The entire experiment is repeated for 100 macro-replications and the results are summarized in Table 1. Table 1 shows the ERMSEs averaged over 100 macro-replications; the values in parentheses are the corresponding standard errors. We observe that at any given level of intrinsic variability (i.e., any value of γ), SKQ dominates the other two methods whereas UC performs the worst, in terms of predicting individual response surfaces corresponding to $z = 1,2,3$ and the overall predictive performance. This is not surprising since UC is designed for deterministic computer experiments and hence its predictive

Table 1: Summary of the average ERMSEs and standard errors for example 1.

	$\gamma = 1$			$\gamma = 25$		
	SKQ	SK	UC	SKQ	SK	UC
ERMSE_1	0.100 (0.003)	0.117 (0.004)	0.29 (0.03)	0.42 (0.01)	0.50 (0.02)	6.7 (5.9)
ERMSE_2	0.081 (0.002)	0.131 (0.008)	0.4 (0.1)	0.36 (0.01)	0.53 (0.02)	1.1 (0.2)
ERMSE_3	0.100 (0.003)	0.13 (0.01)	7.015 (0.010)	0.44 (0.01)	0.53 (0.02)	7.09 (0.10)
ERMSE	0.096 (0.002)	0.132 (0.007)	0.37 (0.07)	0.416 (0.008)	0.54 (0.01)	4.4 (3.4)
	$\gamma = 50$			$\gamma = 100$		
	SKQ	SK	UC	SKQ	SK	UC
ERMSE_1	0.55 (0.02)	0.63 (0.02)	2.2 (0.7)	0.74 (0.02)	0.97 (0.05)	2.4 (0.7)
ERMSE_2	0.49 (0.01)	0.71 (0.02)	1.9 (0.5)	0.64 (0.02)	1.06 (0.06)	2.2 (0.4)
ERMSE_3	0.57 (0.01)	0.71 (0.02)	7.6 (0.3)	0.74 (0.02)	1.14 (0.07)	7.22 (0.07)
ERMSE	0.55 (0.01)	0.70 (0.01)	2.3 (0.5)	0.72 (0.01)	1.13 (0.05)	2.4 (0.4)

performance is expected to be more susceptible to the impact of stochastic simulation errors. The value of γ controls the impact of stochastic simulation errors. We see that the resulting ERMSEs for all three methods increase as γ increases; in particular, the performance of UC deteriorates considerably as γ increases from 1 to 25. On the other hand, on top of the comparable abilities of SKQ and of SK in handling the impact of stochastic simulation errors, it is worthwhile pointing out that the advantage of SKQ is achieved by employing the correlation structure for qualitative factors which makes it possible to capture information across the three closely related response surfaces— even the negative correlations (i.e., $\mathsf{Y}(x,1), \mathsf{Y}(x,3)$ vs. $\mathsf{Y}(x,2)$) are handled appropriately and used effectively—to build metamodels for all three response surfaces simultaneously for prediction.

4.2 Example 2: An (s,S) Inventory System with Holding Cost and Shortage Cost Constraints

In this section, through a simple example of a periodic review (s,S) inventory system, we emphasize the significance of exploiting information from related response surfaces to build a metamodel for better prediction. This example also demonstrates the potential of employing stochastic kriging with qualitative factors to do constrained optimization based on the predicted response surfaces over the design space.

The inventory system is assumed to have random demands, random lead times, full backlogging, and linear ordering, holding and shortage costs. The scenario considered here is the same as discussed in Section 3.1 of Biles et al. (2007), from which much of this example is constructed. The goal is to find the optimal ordering policy $\mathbf{x} = (s,S)^\top$ (in units) that minimizes the expected total costs per day over a simulation period of 120 days subject to constraints on the shortage costs and inventory holding costs per day. Specifically, let $y_1(\mathbf{x})$, $y_2(\mathbf{x})$, $y_3(\mathbf{x})$ denote the expected total cost, the expected holding cost and the expected shortage cost, respectively. Then the problem under consideration can be written as

$$\begin{aligned} \min_{\mathbf{x}=(s,S)^\top} \quad & y_1(\mathbf{x}) \\ \text{subject to} \quad & y_2(\mathbf{x}) \le 25 \\ & y_3(\mathbf{x}) \le 10 \\ & \mathbf{x} \ge \mathbf{0} \ . \end{aligned} \tag{10}$$

Let W_i be the inventory position (on-hand inventory plus outstanding orders minus backlogs on day i). The (s,S) inventory system works as follows. Upon a daily inventory review, if W_i is found below s units, an

order of $(S-W_i)$ units will be made and a fixed ordering cost \$32 per order plus \$3 per unit purchasing cost will be incurred. The order lead time is uniformly distributed between one-half day and one day. The inventory holding cost is \$1 per unit per day and the shortage cost is \$5 per unit per day. The demands arrive as a Poisson process at a rate of 10 customers per day, and the number of units demanded per customer follows a discrete distribution between 1 unit to 4 units whose c.d.f. follows as $\{0.17,1;0.5,2;0.83,3;1,4\}$.

Experiment design. Our approach to solving this constrained optimization problem is by building metamodels for the three underlying response surfaces, namely, y_1, y_2 and y_3. The design space for $\mathbf{x}^\top=(s,S)$ is $s\in[10,40]$ and $S\in[40,90]$. Simulations are conducted at the same set of 20 LHS design points as was used by Biles et al. (2007); and n simulation replications are conducted at each pair of (s,S) with $n\in\{5,50\}$. We are interested in comparing SKQ against SK and the kriging method used in Biles et al. (2007) (denoted by K), in terms of their prediction performances at selected check-points and their abilities of using the resulting predicted response surfaces to do optimization. Notice that the latter two methods SK and K build metamodels for the three surfaces separately without considering the interplay; moreover, K does not take into account the intrinsic variability due to simulation errors through metamodel estimation to prediction.

In applying SKQ, we use an extra three-level qualitative factor z to denote which response surface is under consideration. Specifically, at design point $\mathbf{w}$ for SKQ, i.e., $\mathbf{w}=(\mathbf{x}^\top,z)^\top$ with $\mathbf{x}^\top=(s,S)$, the simulation outputs on the jth replication are modeled as follows,

$$\mathscr{Y}_j(\mathbf{w})=\begin{cases}\text{estimated total cost at } (s,S) \text{ on } j\text{th replication}, & \text{if } z=1,\\ \text{estimated holding cost at } (s,S) \text{ on } j\text{th replication}, & \text{if } z=2,\\ \text{estimated shortage cost at } (s,S) \text{ on } j\text{th replication}, & \text{if } z=3.\end{cases}$$

A grid of 25 equally spaced check-points (s,S) in $[20,40]\times[40,80]$ are used to evaluate the performances of SKQ and SK in predicting y_1,y_2,y_3. True values of the responses at the 25 check-points are approximated by simulating an Arena-based model for 3000 replications, with which the absolute relative prediction error (ARPE) at each check-point is calculated. We examine the quartiles of the ARPEs over the 25 check-points, i.e., the 25th, 50th, and 75th percentiles of the resulting ARPEs.

Results. The predicted response surfaces of y_1, y_2 and y_3 by SKQ and SK based on 5 simulation replications at the design points are shown in Figures 2–4, respectively. We find that the surfaces given by these two methods differ considerably; to quantify the differences, the resulting ARPE quartiles in accordance with $n=5$ and 50 simulation replications applied at each design point are summarized in Table 2. For both SKQ and SK, we observe that increasing the number of replications from 5 to 50 reduces ARPEs. Comparisons of the resulting ARPEs indicate that the response surface for the expected shortage cost is the most difficult to predict well. This manifests one of the challenges facing researchers and practitioners when applying metamodeling techniques to do constrained optimization, a crucial step if not more important than obtaining a good predicted response surface for the objective, is to approximate response surfaces for constraints sufficiently well. Furthermore, it is clear from Table 2 that SKQ outperforms SK in predicting all three surfaces. For this particular example, since the surface of expected total cost, and the surfaces for the expected holding and shortage costs must be closely related to one another, it is not surprising to see that SKQ performs better than SK by exploiting the information available for all three surfaces.

As to optimization, we face the same challenges as Biles et al. (2007) did, that is, to get an approximated optimal solution from the predicted response surface of the objective and at the same time check its feasibility through the predicted surfaces of the two constraints; in addition, an optimal solution (s^*,S^*) is assumed to be integer-valued and so a localized search is necessary. Biles et al. (2007) claim that the true optimal solution found by Excel Solver is $(s,S)=(25,63)$ with a total cost of \$118.47; and the two constraints in (10) are not binding at that solution. We did a localized simulation search for $s\in\{24,25,26\}$ and $S\in\{62,63,64\}$ to approximate the true values of y_1,y_2,y_3 at each pair of (s,S) through simulating 3000 replications with Arena. Our approximated optimal solution with Arena is $(s,S)=(25,62)$ leading to a

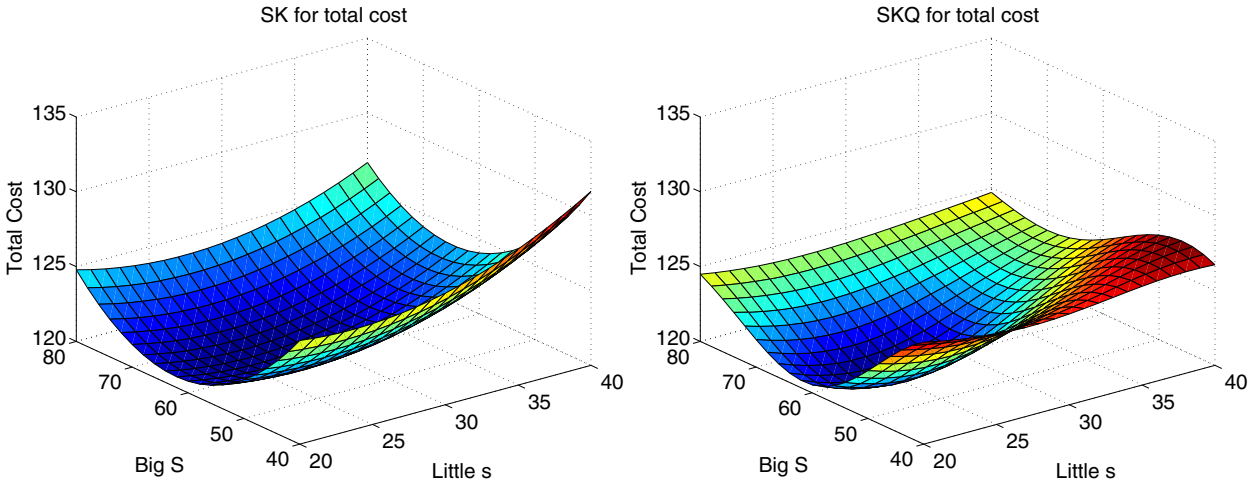

Figure 2: The predicted surfaces of expected total cost by SK (left) and SKQ (right).

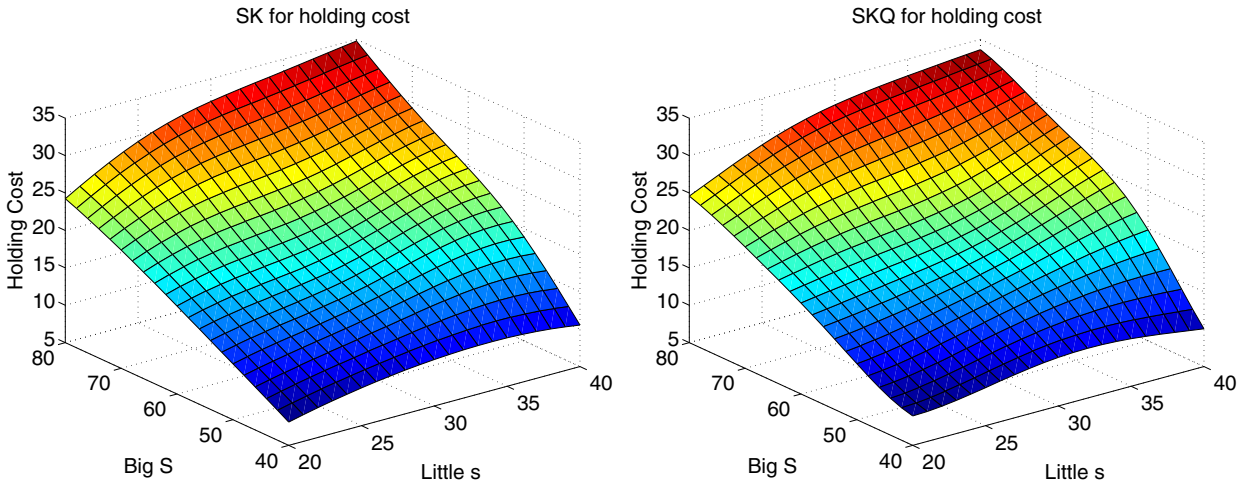

Figure 3: The predicted surfaces of expected holding cost by SK (left) and SKQ (right).

Table 2: Summary of the ARPE quartiles obtained with 5 and 50 simulation replications for example 2.

		y_1			y_2			y_3		
n	Percentiles	25th	50th	75th	25th	50th	75th	25th	50th	75th
5	SKQ	0.006	0.013	0.017	0.011	0.022	0.034	0.059	0.082	0.202
	SK	0.012	0.015	0.018	0.015	0.022	0.036	0.064	0.106	0.367
50	SKQ	0.001	0.002	0.004	0.005	0.008	0.017	0.023	0.042	0.150
	SK	0.001	0.003	0.006	0.005	0.011	0.029	0.023	0.046	0.175

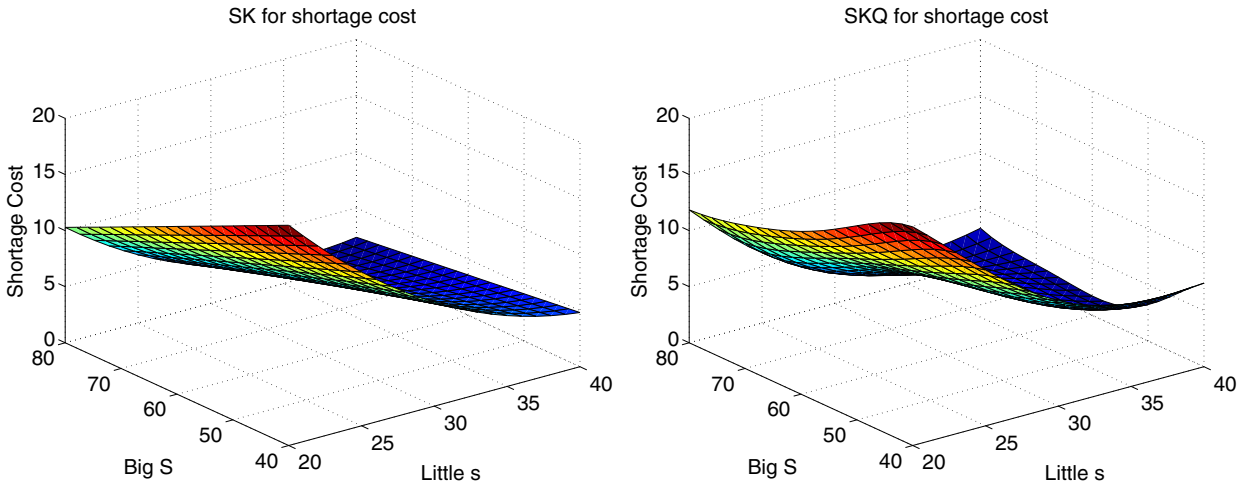

Figure 4: The predicted surfaces of shortage cost by SK (left) and SKQ (right).

total cost of \$118.08. In fact, the local search with intensive simulations reveals that no constraints are binding and the approximated total costs are very close on this 3×3 grid, that is, the surface of the objective is flat in this neighborhood.

Having a sense of where the true optimal solution lies, we are ready to compare the competing methods under the same conditions, i.e., use 5 replications at the same set of 20 LHS design points. The optimal solution given by SK is $(s,S) = (26,62)$ with an estimated total cost of \$120.15; SKQ gives an estimated total cost of \$119.36 at $(s,S) = (25,62)$, which coincides with the optimal solution found by Arena. We see that both SK and SKQ give solutions that are very close to the approximated optimal solution, despite that their predicted total costs are a little off. In sharp contrast, the optimal solution found by K in Biles et al. (2007) is $(s,S) = (27,60)$ with an estimated total cost of \$118.56. This solution is neither close to the optimal solution $(s,S) = (25,63)$ suggested by Excel Solver, nor is it close to the optimal solution $(s,S) = (25,63)$ found by Arena. In this example, K, like SK, does not utilize the correlations existing among the response surfaces of the objective and the two constraints. We remind ourselves that for the purpose of optimization, it is more important to find a solution closer to the true optimal rather than to obtain a more accurate objective value at a suboptimal solution. Therefore, we consider that SKQ has great potentials to be explored as a useful tool for simulation optimization.

5 CONCLUSIONS

In this paper, we developed a new metamodeling method called stochastic kriging with qualitative factors (SKQ) for stochastic simulations involving both qualitative and quantitative decision variables, as an extension of the standard stochastic kriging methodology. We exhibited the main ingredients in SKQ metamodel construction, estimation and prediction. In particular, the recipe for constructing valid spatial correlation functions for handling correlations across levels of qualitative factors is examined. Two examples are shown to demonstrate the potential of SKQ in exploiting information from correlated response surfaces and its ability of providing adequate approximations to multiple response surfaces simultaneously, under the influence of stochastic simulation errors. The impact of stochastic simulation errors and the joint modeling of simulated response surfaces having varying noise levels on the performance of SKQ, as well as efficient experimental designs for SKQ for prediction and simulation optimization are some of the potential research topics await to be explored in the future.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Qiang Zhou from the City University of Hong Kong for his helpful comments and suggestions. This work was supported by National Science Foundation Grant No. CBET-1065931.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic kriging for simulation metamodeling". *Operations Research* 58:371–382.

Biles, W., J. Kleijnen, W. C. M. van Beers, and I. van Nieuwenhuyse. 2007. "Kriging metamodeling in constrained simulation optimization: an explorative study". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 355–362. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Chen, X., B. E. Ankenman, and B. L. Nelson. 2012. "The effects of common random numbers on stochastic kriging metamodels". *ACM Transactions on Modeling and Computer Simulation* 22:7/1–7/20.

Han, G., T. J. Santner, W. I. Notz, and D. L. Bartel. 2009. "Prediction for computer experiments having quantitative and qualitative input variables". *Technometrics* 51:278–288.

Qian, Z. G. P., H. Q. Wu, and C. F. J. Wu. 2008. "Gaussian process models for computer experiments with qualitative and quantitative factors". *Technometrics* 50:192–204.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. "Design and analysis of computer experiments". *Statistical Science* 4:409–423.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. NY: Springer.

Zhou, Q., P. Z. G. Qian, and S. Zhou. 2011. "A simple approach to emulation for computer models with qualitative and quantitative factors". *Technometrics* 53:266–273.

AUTHOR BIOGRAPHIES

XI CHEN is an Assistant Professor in the Department of Statistical Sciences and Operations Research at Virginia Commonwealth University. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization. Her email address is xchen4@vcu.edu and her web page is http://www.people.vcu.edu/~xchen4/.

KAI WANG is a Ph.D. student in the Department of Industrial and Management Systems Engineering at West Virginia University. His email address is kaiwang1987@gmail.com.

FENG YANG is an Assistant Professor in the Department of Industrial and Management Systems Engineering at West Virginia University. Her email address and web page are, respectively, Feng.Yang@mail.wvu.edu and http://www2.statler.wvu.edu/~yang/.