OPERATIONS RESEARCH
Vol. 58, No. 2, March–April 2010, pp. 371–382
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5802 | 0371

DOI 10.1287/opre.1090.0754
© 2010 INFORMS

Stochastic Kriging for Simulation Metamodeling

Bruce Ankenman, Barry L. Nelson, Jeremy Staum
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208
{ankenman@northwestern.edu, nelsonb@northwestern.edu, j-staum@northwestern.edu}

We extend the basic theory of kriging, as applied to the design and analysis of deterministic computer experiments, to the stochastic simulation setting. Our goal is to provide flexible, interpolation-based metamodels of simulation output performance measures as functions of the controllable design or decision variables, or uncontrollable environmental variables. To accomplish this, we characterize both the intrinsic uncertainty inherent in a stochastic simulation and the extrinsic uncertainty about the unknown response surface. We use tractable examples to demonstrate why it is critical to characterize both types of uncertainty, derive general results for experiment design and analysis, and present a numerical example that illustrates the stochastic kriging method.

Subject classifications: simulation: design of experiments, statistical analysis.
Area of review: Simulation.
History: Received May 2008; revisions received October 2008, November 2008; accepted November 2008. Published online in *Articles in Advance* December 8, 2009.

1. Introduction

Discrete-event simulation is a general-purpose tool for analyzing dynamic, stochastic systems. Virtually any level of detail can be modeled and any performance measure estimated, which explains simulation's popularity. However, simulation models are often tedious to build, need substantial data for input modeling, and require significant time to run, particularly when there are many alternatives to evaluate. The decision to build and use a simulation model of a large-scale, complex system often represents a nontrivial investment of time and money.

The objective of the methodology described in this paper is to get more benefit from a simulation investment. The specific context we have in mind is when time to exercise the simulation model in advance of the decision making it will support is relatively plentiful, but decision-making or decision-maker time is relatively scarce or expensive. Therefore, rather than executing a simulation run whenever a "what if" question is posed, or trying to anticipate every scenario of interest in advance, we use the simulation to "map" the performance response surfaces of interest as functions of the controllable design or decision variables, or uncontrollable environmental variables. Ideally, these response surface maps provide the fidelity of the full simulation model with the ease of use of, say, a spreadsheet model.

The motivation for this work is our experience with two industries that build large-scale simulation models: automobile and semiconductor manufacturing. In the automobile application the response was throughput and the controllable design variables included machine capacity, process cycle times, mean time to failure, and mean time to repair, which were controllable through choice of technology. In the semiconductor application the response was start-to-finish product cycle time, and the design variables were product start rates. A key similarity in both settings is that the full simulation model was too slow and clumsy to support the way that decisions were actually made by decision makers trading off performance against less quantifiable objectives.

Using simulation to construct metamodels (models of the simulation model) is not new (see Barton and Meckesheimer 2006 for a review). Starting with classical response-surface modeling in statistics (e.g., Myers and Montgomery 2002), simulation researchers have adapted experiment designs for linear regression models to account for dependence within a replication for steady-state simulations (e.g., Law and Kelton 2000); to permit the use of common random numbers (CRN) and antithetic variates across design points (e.g., Schruben and Margolin 1978; Nozari et al. 1987; Tew and Wilson 1992, 1994); and to compensate for the strong relationship between response variance and customer load in queueing simulations (e.g., Cheng and Kleijnen 1998, Yang et al. 2007). However, linear regression models (that are usually polynomials in the design variables and linear in their unknown coefficients) tend to fit well locally, but do not provide the sort of robust global maps we desire. Nonlinear models based on queueing theory work very well for queueing simulations, but require domain knowledge of the problem context and specialized fitting algorithms.

We are interested in more general-purpose approaches that assume less structure than linear or queueing-specific nonlinear models; that tend to be more resistant to overfitting than general interpolators (e.g., neural networks, see for instance, Sabuncuoglu and Touhami 2002); that facilitate sequential, adaptive experiment designs rather than fixed, a priori designs; and that can provide statistical inference about when a good fit is obtained. We also want to account for the reality that the simulation output is stochastic, with variance that usually changes significantly across the design space.

To satisfy these requirements, we extend the kriging methodology that is popular, and has been highly successful, in the design and analysis of (deterministic) computer experiments (DACE). DACE methodology is particularly well suited for systematically reducing uncertainty about the unknown response surface as experiments (computer runs at different design settings) are performed, and leads to interpolation-based models. Our central contribution is to fully account for the sampling variability that is inherent to a stochastic simulation. We show that correctly accounting for both sampling and response-surface uncertainty has an impact on experiment design, response-surface estimation, and inference.

In the next section we describe our extended metamodel under the special case that all model parameters are known; this setting allows us to demonstrate why the extension is critical without cluttering the discussion with estimation issues, which are resolved in §3. A numerical illustration and conclusions close the paper in §§4 and 5, respectively.

2. The Metamodel

We describe our approach by refining a sequence of models. We are interested in modeling an unknown performance-measure surface (or surfaces) $y(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top$ is a vector of design variables and $y(\mathbf{x})$ is a deterministic function of $\mathbf{x}$. For instance, in a semiconductor fabrication simulation $\mathbf{x}$ might represent the release rates of d products, and y could be the steady-state mean cycle time of product 1 (however, y need not be a mean).

The classical linear regression approach is to assume that the *observed* response obtained from the jth simulation replication at $\mathbf{x}$ is described by the model

$$Y_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \varepsilon_j(\mathbf{x}), \tag{1}$$

where $\mathbf{f}(\mathbf{x})$ is a vector of known functions of $\mathbf{x}$, $\boldsymbol{\beta}$ is a vector of unknown parameters of compatible dimension, and $\varepsilon_j(\mathbf{x})$ has mean 0 and represents the sampling variability inherent in a stochastic simulation. The distribution of $\varepsilon_j(\mathbf{x})$, and in particular its variance, may depend on $\mathbf{x}$, although this dependence is often ignored. We refer to ε as *intrinsic* uncertainty, because it comes from the nature of the stochastic simulation itself. An experiment design specifies settings of $\mathbf{x}$ at which to observe $Y(\mathbf{x})$, and the number of replications to obtain at each $\mathbf{x}$. In this paper we primarily address the replication setting (as opposed to the single-run experiment design sometimes used in steady-state simulation).

Now consider the following thought experiment: Suppose that the response $y(\mathbf{x})$ could be observed *without noise*, but we are still interested in developing a metamodel after observing $y(\mathbf{x})$ at a few design points $\mathbf{x}$. This problem is treated in the DACE literature (Kennedy and O'Hagan 2000, Sacks et al. 1989, Stein 1999, Santner et al. 2003). A remarkably successful approach is to cast this deterministic problem into a statistical framework by representing the unknown response surface as

$$\mathsf{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}), \tag{2}$$

where M is a realization of a mean 0 *random field*; that is, we think of M as being randomly sampled from a space of functions mapping $\Re^d \to \Re$. The functions in this space are assumed to exhibit *spatial correlation*, which means that values $\mathsf{M}(\mathbf{x})$ and $\mathsf{M}(\mathbf{x}')$ will tend to be similar if $\mathbf{x}$ and $\mathbf{x}'$ are close to each other in space. We refer to the stochastic nature of M as *extrinsic* uncertainty, because it is imposed on the problem (not intrinsic to it) to aid in developing a metamodel. This paradigm embeds a deterministic problem into a probabilistic framework so that statistical concepts such as mean squared error (MSE) of estimation can be brought to bear. Statistical inference about $\mathsf{Y}(\mathbf{x})$ at values of $\mathbf{x}$ not simulated can aid experiment design and provide estimates of the metamodel's precision, a feature we want to exploit.

We argue that the following model is more useful than (1) or (2) for representing a stochastic simulation's output on replication j at design point $\mathbf{x}$:

$$\mathcal{Y}_j(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}). \tag{3}$$

The intrinsic noise $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ at a design point $\mathbf{x}$ is naturally independent and identically distributed across replications, but we allow the possibility that $\mathsf{V}(\mathbf{x}) \equiv \mathrm{Var}[\varepsilon(\mathbf{x})]$ is not constant and that $\mathrm{Corr}[\varepsilon_j(\mathbf{x}), \varepsilon_j(\mathbf{x}')] > 0$ to model the effect of CRN. (The intent of CRN is to reduce the variance of estimated differences through inducing positive correlation across design points by driving their simulations with the same sequence of pseudorandom numbers; see, for instance, Law and Kelton 2000.) Later we propose simultaneously modeling M and V, which is a central contribution of this paper.

In our setting, an experiment design consists of pairs $(\mathbf{x}_i, n_i)$, $i = 1, 2, \ldots, k$, where n_i is the number of simulation replications taken at design setting $\mathbf{x}_i$. Let the sample mean at $\mathbf{x}_i$ be

$$\bar{\mathcal{Y}}(\mathbf{x}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{Y}_j(\mathbf{x}_i) \tag{4}$$

and let $\bar{\mathcal{Y}} = (\bar{\mathcal{Y}}(\mathbf{x}_1), \bar{\mathcal{Y}}(\mathbf{x}_2), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))^\top$.

We want a metamodel that predicts the response $\mathsf{Y}(\mathbf{x}_0) \equiv \mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_0)$ at *any* $\mathbf{x}_0$, simulated or not. Until further notice, we only consider the case $\mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta} = \beta_0$ (that is, just a constant term representing the overall surface mean), because this model has tended to be the most useful in practice for DACE.

As is typical in spatial correlation models, we consider linear predictors of the form

$$\lambda_0(\mathbf{x}_0) + \boldsymbol{\lambda}(\mathbf{x}_0)^\top \bar{\mathcal{Y}}, \tag{5}$$

where $\lambda_0(\mathbf{x}_0)$ and $\boldsymbol{\lambda}(\mathbf{x}_0)$ are weights that depend on $\mathbf{x}_0$ and are chosen to give the predictor good properties, such as minimum MSE for predicting $\mathsf{Y}(\mathbf{x}_0) = \beta_0 + \mathsf{M}(\mathbf{x}_0)$. Later, when we make Gaussian assumptions on the intrinsic and extrinsic uncertainty, this form drops out as the best predictor, linear or otherwise.

Let $\Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[\mathsf{M}(\mathbf{x}), \mathsf{M}(\mathbf{x}')]$ be the covariance implied by the extrinsic spatial correlation model, let Σ_{M} be the $k \times k$ covariance matrix across all design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, and let $\Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ be the $k \times 1$ vector $(\mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_1)], \mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_2)], \ldots, \mathrm{Cov}[\mathsf{M}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_k)])^\top$. Also, let Σ_ε be the $k \times k$ covariance matrix implied by the intrinsic noise with (h, i) element $\mathrm{Cov}[\sum_{j=1}^{n_h} \varepsilon_j(\mathbf{x}_h)/n_h, \sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i)/n_i]$ across all design points $\mathbf{x}_h$ and $\mathbf{x}_i$.

To illustrate the key issues, suppose that $\Sigma_{\mathsf{M}}, \Sigma_\varepsilon$, and β_0 are known (clearly, in a real application they need to be estimated, which is a contribution of our research). In the e-companion to this paper, which is available as part of the online version that can be found at http://or.journal.informs.org/, we show that the MSE-optimal predictor of the form (5) is

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \beta_0 + \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top [\Sigma_{\mathsf{M}} + \Sigma_\varepsilon]^{-1} (\bar{\mathcal{Y}} - \beta_0 \mathbf{1}_k), \tag{6}$$

where $\mathbf{1}_k$ is the $k \times 1$ vector of ones. We refer to this predictor as *stochastic kriging*. Notice that the only computationally intensive operation in evaluating (6) is the matrix inversion, which is done once because it is independent of $\mathbf{x}_0$. If there were no intrinsic uncertainty due to simulation, Σ_ε would vanish and (6) would reduce to the standard kriging estimator that matches the data $\bar{\mathcal{Y}}$ at design points, and predicts $\mathsf{Y}(\mathbf{x}_0)$ by a weighted average of $\bar{\mathcal{Y}}$ elsewhere (e.g., Cressie 1993). Equation (6) clearly shows that the presence of intrinsic uncertainty impacts the prediction everywhere on the surface. In the e-companion to this paper, we also show that the optimal MSE is

$$\begin{aligned}\mathrm{MSE}^\star &= \Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \big[\Sigma_{\mathsf{M}} + \Sigma_\varepsilon\big]^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot) \\ &= \big[\Sigma_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \Sigma_{\mathsf{M}}^{-1} \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)\big] \\ &\quad + \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot)^\top \Xi\, \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot),\end{aligned} \tag{7}$$

where Ξ is a positive definite matrix that depends on Σ_ε and Σ_{M}. The term in brackets in (7) is the usual kriging MSE; the additional term is positive, showing that intrinsic uncertainty inflates MSE.

To actually estimate a stochastic kriging metamodel from data, we need $\Sigma_{\mathsf{M}}(\cdot, \cdot)$ to have more structure. In particular, we will assume that M is second-order stationary, meaning that

$$\Sigma_{\mathsf{M}}(\mathbf{x}, \mathbf{x}') = \tau^2 R_{\mathsf{M}}(\mathbf{x} - \mathbf{x}'; \boldsymbol{\theta}), \tag{8}$$

where τ^2 can be interpreted as the variance of $\mathsf{M}(\mathbf{x})$ for all $\mathbf{x}$, and R_{M} is the correlation that depends only on $\mathbf{x} - \mathbf{x}'$ and may be a function of some unknown parameters $\boldsymbol{\theta}$. Further, we will require that $R_{\mathsf{M}}(\mathbf{x} - \mathbf{x}'; \boldsymbol{\theta}) \to 0$ as the distance between $\mathbf{x}$ and $\mathbf{x}'$ goes to infinity, and $\mathbf{R}_{\mathsf{M}}(\mathbf{0}; \boldsymbol{\theta}) = 1$.

Results similar to (6) and (7) have appeared in other contexts, with other interpretations. In the application of kriging to spatial statistics, measurement error can lead to the so-called "nugget effect," which inserts a term that might be represented as $\Sigma_\varepsilon = \sigma^2 \mathbf{I}$; see, for instance, Cressie (1993, Chapter 3). Similarly, O'Hagan and Forster (2004, Chapter 13) describe a Bayesian nonparametric regression setting in which the regression function has a Gaussian process prior and the observations have measurement error, leading to expressions analogous to (6) and (7). Also, in the study of variance components, where the goal might be to predict a random effect such as the IQ of a person drawn from a population, similar expressions arise when the experiment consists of multiple subjects, and multiple tests per subject. See, for instance, Searle et al. (1992, Chapter 7).

Use of kriging for metamodeling in stochastic simulation was first mentioned by Mitchell and Morris (1992), but has only been explored in depth by Kleijnen and his collaborators; the papers most closely related to our work are van Beers and Kleijnen (2003) and Kleijnen and van Beers (2005) (see also Biles et al. 2007 and van Beers and Kleijnen 2008). The central idea in these papers is to first model out any trend using least-squares or generalized least-squares techniques, and then to apply kriging to some form of standardized residuals. They do not incorporate a model of the intrinsic uncertainty, which means that they cannot be used for the sort of adaptive design we desire, which jointly considers the placement of design points and simulation effort. To illustrate the insights gained from our approach, we examine two tractable examples in detail.

2.1. A Two-Point Problem

Consider the case of $k = 2$ design points $\mathbf{x}_1$ and $\mathbf{x}_2$ with equal numbers of replications $n_1 = n_2 = n$. Suppose that

$$\Sigma_{\mathsf{M}} = \tau^2 \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_{\mathsf{M}}(\mathbf{x}_0, \cdot) = \tau^2 \begin{pmatrix} r_0 \\ r_0 \end{pmatrix}.$$

The term $\tau^2 > 0$ represents the extrinsic variance of M, r_{12} is the extrinsic correlation between $\mathsf{M}(\mathbf{x}_1)$ and $\mathsf{M}(\mathbf{x}_2)$, and r_0 is the extrinsic correlation between the point to be predicted $\mathsf{Y}(\mathbf{x}_0)$ and each of the design points (these usually would not be equal). Typically, we expect r_{12} and r_0 to be positive.

For the intrinsic uncertainty due to sampling at a design point, suppose

$$\Sigma_\varepsilon = \frac{\mathsf{V}}{n} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where in this example the variance at the design points is a common $\mathsf{V} > 0$, and $-1 \leqslant \rho \leqslant 1$ represents intrinsic dependence between the design points; for instance, we would expect $\rho > 0$ if we used CRN. Substituting these into (6)–(7), the MSE-optimal predictor of $\mathsf{Y}(\mathbf{x}_0)$ is

$$\widehat{\mathsf{Y}}(\mathbf{x}_0) = \beta_0 + \frac{2\tau^2 r_0}{(1+r_{12})\tau^2 + (1+\rho)\mathsf{V}/n} \cdot \left(\frac{\bar{\mathcal{Y}}(\mathbf{x}_1) + \bar{\mathcal{Y}}(\mathbf{x}_2)}{2} - \beta_0 \right), \qquad (9)$$

with MSE

$$\mathrm{MSE}^\star = \tau^2 \left(1 - \frac{2\tau^2 r_0^2}{(1+r_{12})\tau^2 + (1+\rho)\mathsf{V}/n} \right). \qquad (10)$$

Equation (9) shows that stochastic kriging is a bit like a control-variate estimator (e.g., Nelson 1990), where a correction term is applied to the mean based on the deviation of the observed responses from their expectations and the strength of the correlation (r_0) between the design points and the response to be predicted.

The MSE (10) is even more revealing: MSE is decreasing in r_0^2, meaning the stronger the correlation between the design points and the response at $\mathbf{x}_0$, the smaller the MSE because the design points provide more information. However, MSE is increasing in r_{12}, because the more correlated the design points themselves are, the less additional information they provide. Intrinsic uncertainty, V, also increases MSE, but can be reduced by increasing the sample size n. Most interesting is that the assumed impact of CRN, which is to make $\rho > 0$, *increases* MSE relative to independent sampling. This may seem surprising, because in standard linear regression models such as (1) the impact of CRN is to reduce the variance of the slope coefficients. However, the stochastic kriging predictor is a weighted average of the outcomes from the design points, and CRN inflates the variance of averages. In fact, (10) shows that antithetic variates (e.g., Law and Kelton 2000), which try to induce $\rho < 0$, would reduce MSE. In the e-companion to this paper, we show that the detrimental effect of CRN persists when there are $k > 2$ design points.

There are two messages in this example: (i) *In stochastic kriging there is an important interplay between the placement of design points* (*through their extrinsic correlation with each other*) *and the simulation effort at the design points* (*through their intrinsic variance*); *and* (ii) *CRN will not be helpful for predicting* $\mathsf{Y}(\mathbf{x})$ *in general.* In the next example we examine message (i) more deeply.

2.2. Noiseless M/M/1 Queue

In this example we move a step closer to realistic system simulation problems to illustrate the importance for experiment design of having a model for the intrinsic uncertainty (specifically, $\mathsf{V}(\mathbf{x})$ the variance of $\varepsilon(\mathbf{x})$).

Let $y(x)$ be the steady-state expected number of customers in an M/M/1 queue with service rate 1 and arrival rate $0 \leqslant x < 1$. Then it is well known that $y(x) = x/(1-x)$. This is the surface we are trying to model.

Let $N_t(x)$ be the observed number of customers in this M/M/1 system at time t. If we were trying to estimate $y(x)$ via simulation, then the natural estimator is

$$\bar{\mathcal{Y}}(x) = t^{-1} \int_0^t N_s(x)\, ds,$$

the average number in the system during t units of simulated time. In this example only, we measure simulation effort by the run length t rather than by the number of replications.[1] For large t, $\mathrm{Var}[\bar{\mathcal{Y}}(x)] \approx \mathsf{V}(x)/t \equiv 2x(1+x)/(t(1-x)^4)$ (Whitt 1989). We use this knowledge to examine the impact of design point placement $\{x_1, x_2, \ldots, x_k\}$ and corresponding effort allocation $\{t_1, t_2, \ldots, t_k\}$ on the stochastic kriging estimator without actually simulating the system. To focus the analysis, we suppose that $\bar{\mathcal{Y}}(x)$ is unbiased for $y(x)$, which would occur if we initialized the simulation in steady state.

To represent the surface $y(x)$ in the form $\beta_0 + \mathsf{M}(x)$, let

$$\beta_0 = (x_U - x_L)^{-1} \int_{x_L}^{x_U} \frac{x}{1-x}\, dx$$

be the mean value of the response function over the interval of interest $[x_L, x_U]$. If we pretend that $y(x)$ is a realization of a stationary random field, then a reasonable stand-in for the extrinsic covariance function is

$$\Sigma_{\mathsf{M}}(x, x') = c(h) = (x_U - x_L)^{-1} \int_{x_L}^{x_U - h} \left(\frac{x}{1-x} - \beta_0 \right) \cdot \left(\frac{x+h}{1-x-h} - \beta_0 \right) dx,$$

where $h = |x - x'|$. Conceptually, $c(\cdot)$ is the limit of the empirically estimated covariance function we would obtain by observing $y(x)$ at an increasingly fine grid of evenly spaced design points x, which should allow us to produce the best-possible representation of $y(x)$ via stochastic kriging.

Given a design $\{(x_i, t_i), i = 1, 2, \ldots, k\}$, we have

$$\Sigma_{\mathsf{M}} = \begin{pmatrix} c(0) & c(|x_1 - x_2|) & \cdots & c(|x_1 - x_k|) \\ c(|x_2 - x_1|) & c(0) & \cdots & c(|x_2 - x_k|) \\ \vdots & \vdots & \ddots & \vdots \\ c(|x_k - x_1|) & c(|x_k - x_2|) & \cdots & c(0) \end{pmatrix} \qquad (11)$$

and

$$\Sigma_{\mathsf{M}}(x_0, \cdot)^\top = (c(|x_0 - x_1|), c(|x_0 - x_2|), \ldots, c(|x_0 - x_k|)), \qquad (12)$$

whereas

$$\Sigma_\varepsilon = \begin{pmatrix} \mathsf{V}(x_1)/t_1 & 0 & \cdots & 0 \\ 0 & \mathsf{V}(x_2)/t_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathsf{V}(x_k)/t_k \end{pmatrix}. \qquad (13)$$

All of these matrices can be computed numerically, so we can evaluate the MSE (7) of the stochastic kriging estimator as a function of x_i and t_i. Because we are interested in global fitting, one reasonable objective suggested by Sacks et al. (1989) is to place design points (and in our case also simulation effort) to minimize the integrated MSE

$$\text{IMSE} = \int_{x_L}^{x_U} \text{MSE}^\star(x)\, dx, \tag{14}$$

where $\text{MSE}^\star(x)$ is the MSE of the optimal predictor at x, as in (7).

Consider the specific case $x_L = 0.5$ and $x_U = 0.95$, where we take x_L and x_U as two of $k = 3$ design points on which we can spend $t = 10,000$ units of simulation effort. If we allocate the simulation effort in units of $1,000$, and can place the third design point at one of $x = 0.55, 0.6, 0.65, \ldots, 0.85$, then the design that minimizes IMSE for stochastic kriging is $(x, t) = (0.5, 1,000)$, $(0.65, 1,000)$, and $(0.95, 8,000)$ with $\text{IMSE} = 4.70$.

Standard kriging—by which we mean ignoring intrinsic uncertainty—finds the optimal design points to be $x = 0.5, 0.8$, and 0.95 and (incorrectly) estimates the IMSE to be 4.27. The actual IMSE, accounting for sampling variability, is 4.93 if we allocate the effort equally among these three design points, and it only drops to 4.91 if we allocate optimally given these design points. This illustrates the need to account for intrinsic uncertainty in design, and that "design" must include both the placement of design points and the allocation of simulation effort. In §3.3 we provide one method to obtain approximately IMSE-optimal designs for stochastic kriging.

3. Parameter Estimation

To actually apply stochastic kriging for simulation metamodeling, a method for estimating the unknown parameters is required. The DACE literature contains several methods and refinements when there is only extrinsic uncertainty; see, for instance, Santner et al. (2003) and Fang et al. (2006). Here we focus on extending the most well-known method—maximum likelihood—to allow for intrinsic uncertainty.

Recall that our model for the simulation output is

$$\mathcal{Y}_j(\mathbf{x}) = \beta_0 + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}).$$

We now adopt the following.

Assumption 1. *The random field* M *is a stationary Gaussian random field, and* $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots$ *are i.i.d.* $\mathsf{N}(0, \mathsf{V}(\mathbf{x}_i))$, *independent of* $\varepsilon_j(\mathbf{x}_h)$ *for all* j *and* $h \neq i$ (*i.e., no CRN*), *and independent of* M.

That M is a stationary Gaussian random field is a standard assumption in DACE. We refer the reader to, for instance, Santner et al. (2003, §2.3.2) for technical details, but in brief this assumption implies that for any finite collection of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ the random vector $(\mathsf{M}(\mathbf{x}_1), \mathsf{M}(\mathbf{x}_2), \ldots, \mathsf{M}(\mathbf{x}_k))^\top$ has a multivariate normal distribution with constant marginal mean 0, variance $\tau^2 > 0$, and positive definite correlation matrix $\mathbf{R}_\mathsf{M}$ such that $\text{Corr}(\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_h))$ depends only on $\mathbf{x}_i - \mathbf{x}_h$. The normality of $\varepsilon_j(\mathbf{x})$ could be anticipated if, for instance, the output of each replication was itself the average of a large number of more-basic random variables (e.g., the average of hundreds of individual product cycle times in the semiconductor fabrication example).

Under Assumption 1, $(\mathsf{Y}(\mathbf{x}_0), \bar{\mathcal{Y}}(\mathbf{x}_1), \ldots, \bar{\mathcal{Y}}(\mathbf{x}_k))$ is multivariate normal (see the e-companion to this paper), and the stochastic kriging predictor (6) is the conditional expectation of $\mathsf{Y}(\mathbf{x}_0)$ given $\bar{\mathcal{Y}}$, making it the minimum MSE predictor (Santner et al. 2003, Theorem 3.2.1).

We begin by assessing the impact of estimating the intrinsic variance Σ_ε, then derive the maximum-likelihood estimators given Σ_ε, and conclude by addressing experiment design.

3.1. Estimating the Intrinsic Variance

In this section we confront the fact that V is typically unknown. In summary, our approach is as follows:

- Because we are interested in sequential experiment design, we need a model for V. To obtain it, we will assume that V is also represented by a spatial correlation model

$$\mathsf{V}(\mathbf{x}) = \sigma^2 + \mathsf{Z}(\mathbf{x}), \tag{15}$$

where Z is a mean zero stationary random field that is independent of M. Denote the estimated model by $\widehat{\mathsf{V}}(\mathbf{x})$.

- Because $\mathsf{V}(\mathbf{x}_i)$ is not observable, even at the design points, we let

$$\mathcal{S}^2(\mathbf{x}_i) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathcal{Y}_j(\mathbf{x}_i) - \bar{\mathcal{Y}}(\mathbf{x}_i))^2 \tag{16}$$

stand in for it. Under Assumption 1, $\mathcal{S}^2(\mathbf{x}_i)$ is strongly consistent for $\mathsf{V}(\mathbf{x}_i)$ and has a scaled chi-squared distribution.

- Because we observe $\mathcal{S}^2$, not V, there is extrinsic and intrinsic uncertainty, just as in estimating $\beta_0 + \mathsf{M}$ from $\bar{\mathcal{Y}}$. However, because we are not interested in V except as it impacts our design and analysis, we will ignore the intrinsic uncertainty and fit model (15) using standard kriging as if $\mathcal{S}^2$ had no noise. Therefore, $\widehat{\mathsf{V}}(\mathbf{x}_i) = \mathcal{S}^2(\mathbf{x}_i)$ at design points $\mathbf{x}_i$ because standard kriging interpolates the response at the design points exactly. We will show that the consequences of estimating V in this way are slight as long as the n_i are not too small.
- We do not describe estimation of model (15) from $\mathcal{S}^2(\mathbf{x}_1), \mathcal{S}^2(\mathbf{x}_2), \ldots, \mathcal{S}^2(\mathbf{x}_k)$ here, because no new ideas are introduced. In the numerical illustration in §4 we cite a specific approach.

Our first key result is that estimating Σ_ε in this way introduces no prediction bias. The proof can be found in the e-companion to this paper.

Theorem 1. *Let* $\widehat{\Sigma}_\varepsilon = \mathrm{Diag}\{\widehat{\mathsf{V}}(\mathbf{x}_1)/n_1, \widehat{\mathsf{V}}(\mathbf{x}_2)/n_2, \ldots, \widehat{\mathsf{V}}(\mathbf{x}_k)/n_k\}$ *and define*

$$\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) = \beta_0 + \Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\big[\Sigma_{\mathsf{M}} + \widehat{\Sigma}_\varepsilon\big]^{-1}(\overline{\mathcal{Y}} - \beta_0\mathbf{1}_k). \qquad (17)$$

If Assumption 1 *holds, then* $\mathrm{E}[\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)] = 0$.

As a consequence of Theorem 1, our key concern is how much variance inflation occurs when V is estimated. Clearly, if the n_i are large enough, then there is little inflation. But how large do they have to be? To answer this question, we consider another tractable example:

Suppose that

$$\Sigma_{\mathsf{M}} = \tau^2\begin{pmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \vdots & \vdots & \ddots & \vdots \\ r & r & \cdots & 1 \end{pmatrix},$$

$\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot) = \tau^2(r_0, r_0, \ldots, r_0)^\top$ with $r_0, r \geqslant 0$, and $\Sigma_\varepsilon = (\mathsf{V}/n)\mathbf{I}$. This represents a situation in which the extrinsic correlations among the design points are all equal and the design points are equally correlated with the point we wish to predict, which might be (approximately) plausible if the design points are widely separated, say at the extremes of the region of interest, whereas $\mathbf{x}_0$ is central. Note that for the covariance matrix of $(\mathsf{Y}(\mathbf{x}_0), \overline{\mathcal{Y}}(\mathbf{x}_1), \ldots, \overline{\mathcal{Y}}(\mathbf{x}_k))^\top$ to be positive definite, we must have $r_0^2 < 1/k + r(k-1)/k$. The structure of Σ_ε arises because we assume the intrinsic variance is the same across all design points and n replications have been allocated to each of them. Suppose also that we have an estimator $\widehat{\mathsf{V}} \sim \mathsf{V}\chi^2_{n-1}/(n-1)$, meaning that $(n-1)\widehat{\mathsf{V}}/\mathsf{V}$ has a chi-squared distribution. We use a common estimator of the intrinsic variance rather than estimating it at each design point individually to make the example tractable. Finally, let $\gamma = \mathsf{V}/\tau^2$ be the ratio of the intrinsic variance to the extrinsic variance, which is (roughly speaking) a measure of the sampling noise relative to the response-surface variation.

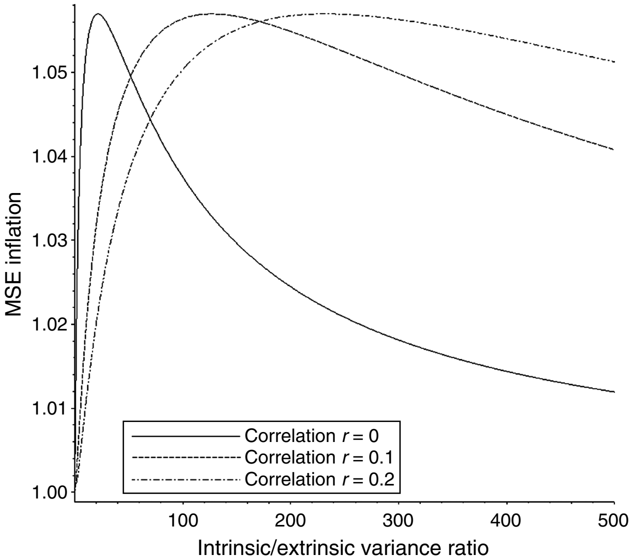
In the e-companion to this paper, we show that the MSE of $\widehat{\mathsf{Y}}(\mathbf{x}_0)$, the stochastic kriging predictor with V known, is

$$\mathrm{MSE}^\star = \tau^2\left(1 - \frac{kr_0^2}{1+(k-1)r+\gamma/n}\right). \qquad (18)$$

On the other hand, the MSE of $\widehat{\widehat{\mathsf{Y}}}(\mathbf{x}_0)$ obtained by substituting $\widehat{\mathsf{V}}$ for V is

$$\mathrm{MSE} = \tau^2\mathrm{E}\left[\left(1 + \frac{(1+(k-1)r+\gamma/n)\,kr_0^2}{(1+(k-1)r+(\gamma/n)(\widehat{\mathsf{V}}/\mathsf{V}))^2} - \frac{2kr_0^2}{(1+(k-1)r+(\gamma/n)(\widehat{\mathsf{V}}/\mathsf{V}))}\right)\right]. \qquad (19)$$

Figure 1. MSE inflation as a function of $\gamma = \mathsf{V}/\tau^2$ when $n = 10$ and correlation r_0 is 95% of its maximum possible value.



We assess the inflation by evaluating the ratio of (19) to (18) numerically. The ratio is largest when n is small and r_0 and r are large, so Figure 1 shows the inflation as a function of $\gamma = \mathsf{V}/\tau^2$ for $n = 10$, $r = 0$, 0.1, 0.2, and r_0 at 95% of the maximum value it can take. Even with this small value of n, the inflation is slight over an extreme range of γ values. As n increases, the inflation vanishes. This suggests that the penalty for estimating V will typically be small, which is further supported by the experiment in §4.

3.2. Maximum-Likelihood Estimation

In this section we derive the maximum-likelihood estimators of $(\beta_0, \tau^2, \boldsymbol{\theta})$, assuming Σ_ε is known. To reduce notation, let $\mathsf{V}_i \equiv \mathsf{V}(\mathbf{x}_i)/n_i$; thus, $\Sigma_\varepsilon = \mathrm{Diag}\{\mathsf{V}_1, \mathsf{V}_2, \ldots, \mathsf{V}_k\}$. Also define $\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})$ to be correlation matrix of M across the design points.

In the e-companion to this paper we show that for a fixed experiment design $\{(\mathbf{x}_i, n_i), i = 1, 2, \ldots, k\}$, and under Assumption 1, the log-likelihood function of $(\beta_0, \tau^2, \boldsymbol{\theta})$ is

$$\begin{aligned} l(\beta_0, \tau^2, \boldsymbol{\theta}) &= -\ln\big[(2\pi)^{k/2}\big] - \tfrac{1}{2}\ln[|\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta}) + \Sigma_\varepsilon|] \\ &\quad - \tfrac{1}{2}(\overline{\mathcal{Y}} - \beta_0\mathbf{1}_k)^\top[\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta}) + \Sigma_\varepsilon]^{-1}(\overline{\mathcal{Y}} - \beta_0\mathbf{1}_k). \qquad (20) \end{aligned}$$

If the Σ_ε terms are removed, then this is the log-likelihood function for kriging when M is a Gaussian random field. We have been intentionally vague about the covariance function $\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})$ because we want the results to be general, but when we apply stochastic kriging later, we will use a standard model from the DACE literature.

Finding the maximum-likelihood estimators requires simultaneously solving

$$\frac{\partial l(\beta_0,\tau^2,\boldsymbol{\theta})}{\partial\beta_0}=0 \qquad \frac{\partial l(\beta_0,\tau^2,\boldsymbol{\theta})}{\partial\tau^2}=0$$

$$\frac{\partial l(\beta_0,\tau^2,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\mathbf{0} \tag{21}$$

for $(\widehat{\beta}_0,\widehat{\tau}^2,\widehat{\boldsymbol{\theta}})$. The primary purpose of this section is to show that when Σ_ε is given, the search to find the MLEs is no more computationally difficult than when Σ_ε is not present, and in fact is more likely to be numerically stable. Complete expressions for (21) are given in the e-companion to this paper.

Let $\Sigma=\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})+\Sigma_\varepsilon$, and let η be generic for any of the unknown parameters β_0, τ^2 or any element of $\boldsymbol{\theta}$. Then, trivially,

$$\frac{\partial\Sigma}{\partial\eta}=\frac{\partial\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})}{\partial\eta}.$$

The elements of this partial derivative matrix are explicit for the typical choices of $\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})$. Applying standard results for matrix calculus, we can show that

$$\frac{\partial|\Sigma|}{\partial\eta}=|\Sigma|\operatorname{trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\eta}\right]=|\Sigma|\operatorname{trace}\left[\Sigma^{-1}\frac{\partial\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})}{\partial\eta}\right] \tag{22}$$

and

$$\frac{\partial\Sigma^{-1}}{\partial\eta}=-\Sigma^{-1}\frac{\partial\Sigma}{\partial\eta}\Sigma^{-1}=-\Sigma^{-1}\frac{\partial\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})}{\partial\eta}\Sigma^{-1}. \tag{23}$$

Thus, the partial derivatives required to solve for the MLEs in (21) are partial derivatives of $\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})$ required in the deterministic DACE case. Of course, the determinant and matrix inverse that must be evaluated are different, namely, $|\Sigma|$ and Σ^{-1} instead of $|\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})|$ and $[\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})]^{-1}$. However, in practice, the correlation matrix $\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})$ can become nearly singular when searching over $(\beta_0,\tau^2,\boldsymbol{\theta})$, causing numerical stability problems in DACE applications of maximum likelihood (Fang et al. 2006). In our case $\Sigma=\tau^2\mathbf{R}_{\mathsf{M}}(\boldsymbol{\theta})+\Sigma_\varepsilon$ is resistant to becoming singular because Σ_ε is not a function of the parameters.

A number of numerical methods can be used to search for the MLEs $(\widehat{\beta}_0,\widehat{\tau}^2,\widehat{\boldsymbol{\theta}})$; see, for instance, Fang et al. (2006). We have had success with using nonlinear optimization routines to maximize the likelihood (20), explicitly including the constraint $\widehat{\tau}^2\geqslant 0$. Any such method will need starting solutions. We have found it helpful to start with moderate values of $\widehat{\beta}_0$, $\widehat{\tau}^2$, and $\widehat{\boldsymbol{\theta}}$, such as initializing $\widehat{\beta}_0$ and $\widehat{\tau}^2$ to the sample average and sample variance of $\bar{\mathscr{Y}}(\mathbf{x}_1),\bar{\mathscr{Y}}(\mathbf{x}_2),\ldots,\bar{\mathscr{Y}}(\mathbf{x}_k)$.

To summarize, given the data $\mathscr{Y}_j(\mathbf{x}_i)$, $j=1,2,\ldots,n_i$, $i=1,2,\ldots,k$, a stochastic kriging metamodel is obtained as follows:

1. Estimate $\widehat{\mathsf{V}}$ as in §3.1 and let $\widehat{\Sigma}_\varepsilon=\operatorname{Diag}\{\widehat{\mathsf{V}}(\mathbf{x}_1)/n_1, \widehat{\mathsf{V}}(\mathbf{x}_2)/n_2,\ldots,\widehat{\mathsf{V}}(\mathbf{x}_k)/n_k\}$ where $\widehat{\mathsf{V}}(\mathbf{x}_i)=\mathscr{S}^2(\mathbf{x}_i)$.
2. Using $\widehat{\Sigma}_\varepsilon$ instead of Σ_ε, maximize the log-likelihood (20) over $(\widehat{\beta}_0,\widehat{\tau}^2,\widehat{\boldsymbol{\theta}})$.
3. Predict $\mathsf{Y}(\mathbf{x}_0)$ by the metamodel

$$\widehat{\mathsf{Y}}(\mathbf{x}_0)=\widehat{\beta}_0+\widehat{\tau}^2\mathbf{R}_{\mathsf{M}}(\mathbf{x}_0,\cdot;\widehat{\boldsymbol{\theta}})^\top[\widehat{\tau}^2\mathbf{R}_{\mathsf{M}}(\widehat{\boldsymbol{\theta}})+\widehat{\Sigma}_\varepsilon]^{-1}(\bar{\mathscr{Y}}-\widehat{\beta}_0\mathbf{1}_k) \tag{24}$$

with MSE estimator

$$\begin{aligned}\widehat{\mathrm{MSE}}(\mathbf{x}_0)&=\widehat{\tau}^2-\widehat{\tau}^4\mathbf{R}_{\mathsf{M}}(\mathbf{x}_0,\cdot;\widehat{\boldsymbol{\theta}})^\top[\widehat{\tau}^2\mathbf{R}_{\mathsf{M}}(\widehat{\boldsymbol{\theta}})+\widehat{\Sigma}_\varepsilon]^{-1}\mathbf{R}_{\mathsf{M}}(\mathbf{x}_0,\cdot;\widehat{\boldsymbol{\theta}})\\&\quad+\boldsymbol{\delta}^\top\boldsymbol{\delta}(\mathbf{1}_k^\top[\widehat{\tau}^2\mathbf{R}_{\mathsf{M}}(\widehat{\boldsymbol{\theta}})+\widehat{\Sigma}_\varepsilon]^{-1}\mathbf{1}_k)^{-1}\end{aligned} \tag{25}$$

where $\boldsymbol{\delta}=1-\mathbf{1}_k^\top[\widehat{\tau}^2\mathbf{R}_{\mathsf{M}}(\widehat{\boldsymbol{\theta}})+\widehat{\Sigma}_\varepsilon]^{-1}\mathbf{R}_{\mathsf{M}}(\mathbf{x}_0,\cdot;\widehat{\boldsymbol{\theta}})\,\widehat{\tau}^2$. The MSE expression is derived in the e-companion to this paper; the last term on the right-hand side of (25) accounts for the variability due to estimating β_0. Both (24) and (25) are plug-in estimators, and therefore (24) is no longer linear in the data.

3.3. Experiment Design

In this section we describe an approach to obtain experiment designs with low IMSE. Our results assume that the extrinsic covariance function $\Sigma_{\mathsf{M}}(\cdot,\cdot)$ and the intrinsic variance function $\mathsf{V}(\cdot)$ are known; later in the section we describe how we might use the results when these functions are estimated.

Let $\mathscr{X}$ be the d-dimensional experiment design space of interest, and suppose that we have k fixed design points $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k$ to which we want to allocate N replications. Let $\mathbf{n}^\top=(n_1,n_2,\ldots,n_k)$. Then our goal is to

$$\text{minimize}\quad \mathrm{IMSE}(\mathbf{n})=\int_{\mathbf{x}_0\in\mathscr{X}}\mathrm{MSE}(\mathbf{x}_0;\mathbf{n})\,d\mathbf{x}_0 \tag{26}$$

$$\text{subject to:}\quad \mathbf{n}^\top\mathbf{1}_k\leqslant N \tag{27}$$

$$n_i \text{ nonnegative integers} \tag{28}$$

where

$$\begin{aligned}\mathrm{MSE}(\mathbf{x}_0;\mathbf{n})&=\Sigma_{\mathsf{M}}(\mathbf{x}_0,\mathbf{x}_0)\\&\qquad-\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)^\top\,[\Sigma_{\mathsf{M}}+\Sigma_\varepsilon(\mathbf{n})]^{-1}\,\Sigma_{\mathsf{M}}(\mathbf{x}_0,\cdot)\end{aligned}$$

and $\Sigma_\varepsilon(\mathbf{n})=\operatorname{Diag}\{\mathsf{V}(\mathbf{x}_1)/n_1,\mathsf{V}(\mathbf{x}_2)/n_2,\ldots,\mathsf{V}(\mathbf{x}_k)/n_k\}$. In words, we minimize the IMSE for the MSE-optimal stochastic kriging estimator as a function of the number of replications allocated to each design point. To obtain an approximate solution to this problem, we relax the integrality constraint (28) and assume only that $n_i\geqslant 0$. Because we will have repeated need of it, let $\Sigma(\mathbf{n})=\Sigma_{\mathsf{M}}+\Sigma_\varepsilon(\mathbf{n})$.

Assuming M is second-order stationary, as in (8), we can let $\Sigma_{\mathsf{M}}(\mathbf{x}_i,\mathbf{x}_0)=\tau^2 r_i(\mathbf{x}_0)$. In the e-companion to this paper we show that the optimal solution $\mathbf{n}^\star$ to (26), with integrality relaxed, satisfies $n_i^\star\propto\sqrt{\mathsf{V}(\mathbf{x}_i)C_i(\mathbf{n}^\star)}$ where

$$C_i(\mathbf{n})=\left[\Sigma(\mathbf{n})^{-1}\mathbf{W}\Sigma(\mathbf{n})^{-1}\right]_{ii}$$

and $\mathbf{W}$ is the $k \times k$ matrix with elements

$$W_{ij} = \int_{\mathbf{x}_0 \in \mathscr{X}} r_i(\mathbf{x}_0) r_j(\mathbf{x}_0)\, d\mathbf{x}_0.$$

To gain some insight into this result, suppose that N is large enough that $\Sigma(\mathbf{n}) \approx \Sigma_{\mathsf{M}}$, so that

$$C_i(\mathbf{n}) \approx C_i = \left[\Sigma_{\mathsf{M}}^{-1} \mathbf{W} \Sigma_{\mathsf{M}}^{-1}\right]_{ii}.$$

Then,

$$n_i^\star \approx N \frac{\sqrt{\mathsf{V}(\mathbf{x}_i) C_i}}{\sum_{j=1}^{k} \sqrt{\mathsf{V}(\mathbf{x}_j) C_j}}. \tag{29}$$

Notice that C_i is a function only of the extrinsic correlation structure, and V is the intrinsic variance. Expression (29) shows how the response surface, as represented by its correlation structure, distorts the allocation of replications from one that is proportional to only the intrinsic standard deviation at the design point; it tends to favor design points that are centrally located because they do more to reduce MSE throughout the design space (notice that W_{ii} will be larger if $\mathbf{x}_i$ is close to more of the design space). This further emphasizes what was illustrated in §2.2: Both intrinsic and extrinsic uncertainty matter in the experiment design.

In practice, neither $\Sigma_{\mathsf{M}}(\cdot,\cdot)$ nor $\mathsf{V}(\cdot)$ are known in advance, and the design points are not given. One way to use these results is via a two-stage design strategy:

1. In Stage 1, select a space-filling design of m predetermined design points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and allocate n_0 replications to each.
2. Fit $\widehat{\mathsf{V}}$ and $\hat{\tau}^2 \mathbf{R}_{\mathsf{M}}(\cdot,\cdot;\hat{\boldsymbol{\theta}})$ as described above.
3. In Stage 2, jointly select $k-m$ additional design points $\mathbf{x}_{m+1}, \ldots, \mathbf{x}_k$ from a larger set and optimally allocate the $N - mn_0$ additional replications among $\mathbf{x}_1, \ldots, \mathbf{x}_k$ to minimize IMSE using $\widehat{\mathsf{V}}$ and $\mathbf{R}_{\mathsf{M}}(\cdot,\cdot;\hat{\boldsymbol{\theta}})$ in place of the true functions. The optimization is facilitated by the fact that

$$\frac{\partial}{\partial n_i} \mathrm{IMSE}(\mathbf{n}) = -\tau^4 \frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} C_i(\mathbf{n}),$$

as we show in the e-companion to this paper, or we can use the approximate formula (29).

To construct a practical, concrete procedure requires making several choices. We discuss some of them in the remainder of this section.

What should the total number of design points, k, be? As in classical two-stage procedures for fixed-width confidence-interval estimation, we could use what we learn in the first stage to choose k large enough to attain an IMSE target. To determine whether a second-stage design $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ attains an IMSE target, we could use an upper bound on the design's IMSE based on an upper bound on MSE at any point $\mathbf{x} \in \mathscr{X}$, such as

$$\mathrm{MSE}(\mathbf{x}) \leqslant \hat{\tau}^2 \left(1 - \max_{i=1,2,\ldots,k} \frac{\hat{\tau}^2 R_{\mathsf{M}}(\mathbf{x}-\mathbf{x}_i;\hat{\boldsymbol{\theta}})}{\hat{\tau}^2 + \mathsf{V}(\mathbf{x}_i)/n_i}\right).$$

This bound follows from Equation (7), and says that (within the framework of plug-in estimation) the MSE of predicting $\mathsf{Y}(\mathbf{x})$ using the design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is no more than the MSE of predicting $\mathsf{Y}(\mathbf{x})$ using only the single design point that is most informative about $\mathsf{Y}(\mathbf{x})$. This leads to upper bounds for IMSE:

$$\begin{aligned}\mathrm{IMSE} &\leqslant \hat{\tau}^2 \left(1 - \min_{\mathbf{x} \in \mathscr{X}} \max_{i=1,2,\ldots,k} \frac{\hat{\tau}^2 R_{\mathsf{M}}(\mathbf{x}-\mathbf{x}_i;\hat{\boldsymbol{\theta}})}{\hat{\tau}^2 + \mathsf{V}(\mathbf{x}_i)/n_i}\right) |\mathscr{X}| \\ &\leqslant \hat{\tau}^2 \left(1 - \frac{\hat{\tau}^2 \min_{\mathbf{x} \in \mathscr{X}, i=1,2,\ldots,k} R_{\mathsf{M}}(\mathbf{x}-\mathbf{x}_i;\hat{\boldsymbol{\theta}})}{\hat{\tau}^2 + \max_{i=1,2,\ldots,k} \mathsf{V}(\mathbf{x}_i)/n_i}\right) |\mathscr{X}|, \end{aligned} \tag{30}$$

where $|\mathscr{X}|$ is the volume of the design space. Expression (30) relates IMSE to the maximum intrinsic uncertainty about the response surface at any design point and the minimum extrinsic correlation between the responses at any point $\mathbf{x} \in \mathscr{X}$ and the nearest design point.

The criterion (30) also helps answer another question: If we start with m initial design points, how should the $k-m$ additional design points $\mathbf{x}_{m+1}, \ldots, \mathbf{x}_k$ be selected? According to (30), the complete design $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ should also be space filling. For example, in a two-dimensional problem, we might take $m = 4$ and use a two-level factorial design in the first stage. Then taking $k = 9$, we could add five design points in a space-filling way, such as completing a three-level factorial design or alternatively using a Latin hypercube design that maximizes the minimum distance between the design points. Stochastic kriging allows us to estimate the resulting IMSE of both designs and choose the best follow-up design.

A second key question is: How should the simulation effort be allocated between the two stages? That is, given the number of design points k and simulation replications N, how should we choose the number of design points m and simulation replications mn_0 to use in the first stage? Choosing n_0 is analogous to the well-known problem of choosing n_0 in classical two-stage procedures for fixed-width confidence-interval estimation or in related ranking-and-selection procedures (e.g., Boesel et al. 2003). If there are too few simulation replications in the first stage, then estimates of $\mathsf{V}(\cdot)$, τ^2, and $\boldsymbol{\theta}$ will be poor, leading to bad decisions about allocating the second-stage budget; if the first-stage computational budget is too large, then the advantage of a two-stage procedure is reduced because there will be less flexibility to allocate the budget adaptively in the second stage.

Unfortunately, just as in ranking and selection, it is difficult to give general guidance about choosing the first-stage budget, other than to say that n_0 should exceed 10 to get useful estimates of intrinsic variance. We can only elaborate on the new issues that are involved in the context of stochastic kriging. These have to do with choosing the number of design points m and k across which mn_0 and N simulation replications are spread. Again, it is difficult to give general guidance because a good allocation depends on the structure of the simulation problem. If the intrinsic variance is low, then it is advantageous to have a large number of design points to fill space thoroughly and reduce extrinsic uncertainty. If the intrinsic variance is high, then the number of design points should not be too large, because when the intrinsic uncertainty about the response surface at design points is too great it will be difficult to estimate τ^2 and $\boldsymbol{\theta}$. The larger $\mathsf{V}(\cdot)$ is compared to τ^2, the fewer design points there should be.

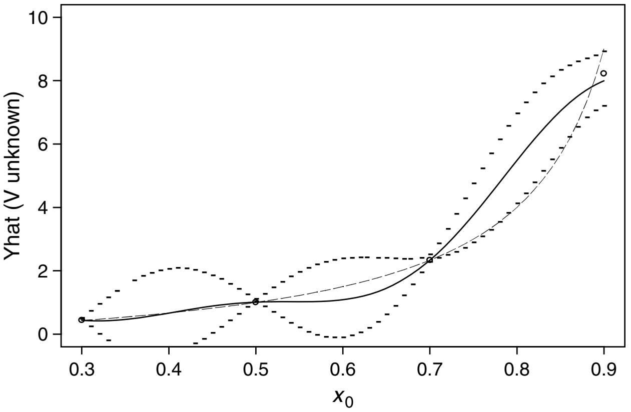
4. Illustration

To illustrate the methodology developed in this paper, we return to the steady-state mean number in an M/M/1 queue, as in §2.2. However, this time we simulate it.

Our purpose in this section is three-fold: To provide some intuition about what the stochastic kriging technique does on a familiar problem; to assess the penalty for estimating the intrinsic covariance matrix Σ_ε; and to evaluate our ability to estimate the error in our metamodel. We do not make direct comparisons to other response-surface modeling techniques, but we note the following: For this particular metamodeling problem, the procedure of Yang et al. (YAN 2007) would undoubtedly be superior to stochastic kriging. YAN is an adaptive procedure designed for queueing performance measures; it fits a nonlinear metamodel that was motivated by known results for the M/M/1 queue. On the other hand, a standard quadratic response-surface model is known to perform poorly for the M/M/1 queue because the polynomial fails to fit $x/(1-x)$ well over a large domain for x (we use $0.3 \leqslant x \leqslant 0.9$ here), and the response variance increases explosively as x increases. We intend stochastic kriging to be used primarily in situations where little is known about the response surface, the same situations in which we would use polynomial regression. Large-scale comparisons with other methods is a subject of future work.

The statistic we record from each replication is the average number of customers in the system from time 0 to T. For the M/M/1 queue we can initialize each replication in steady state by independently sampling the number in the system at time 0 from the steady-state distribution. We keep the run length per replication T the same for all arrival rates x, so that we entirely control intrinsic variance through the number of replications. To assess the penalty for estimating the intrinsic variance, we also apply stochastic kriging using the known variance function $\mathsf{V}(x)/T = 2x(1+x)/(T(1-x)^4)$. We do not employ CRN. For fitting the mean and variance models we assume a Gaussian correlation structure of the form $\mathbf{R}_\mathsf{M}(x_i, x_j; \theta_\mathsf{M}) = \exp(-\theta_\mathsf{M}(x_i - x_j)^2)$ and $\mathbf{R}_\mathsf{V}(x_i, x_j; \theta_\mathsf{V}) = \exp(-\theta_\mathsf{V}(x_i - x_j)^2)$, respectively, with the θs unknown. All of the simulation and fitting of the metamodels was done using our own code written in S-PLUS; fitting was via maximum likelihood.

Figure 2. Fitted via stochastic kriging (solid line) and true (dashed line) expected number in an M/M/1 queue from the first-stage experiment.



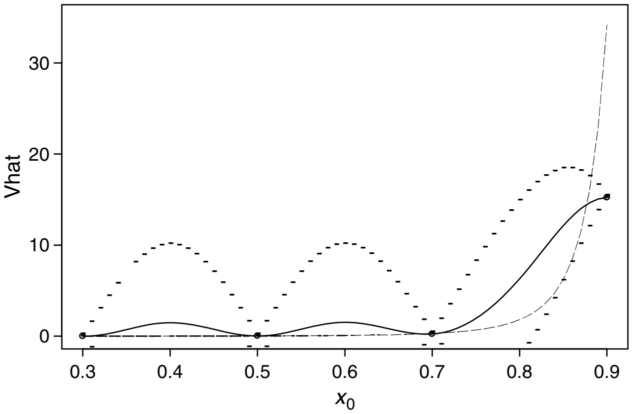
To illustrate stochastic kriging, we consider an experiment that starts with four design points, $x = 0.3$, 0.5, 0.7, 0.9, making 20 replications of length $T = 1{,}000$ time units at each of them (80 replications total). Based on the results, we allocate a total of $N = 500$ replications among these four design points, plus three additional points $x = 0.4$, 0.6, 0.8, using the approximately optimal allocation formula (29), and view the final fit.

Figure 2 shows the results for the mean number in queue metamodel $\widehat{\widehat{\mathsf{Y}}}(x_0)$ from the first-stage experiment. In the plot, a circle represents an estimated response from the simulation (the data points); the solid-line curve is the stochastic kriging metamodel, which is surrounded by $\pm\sqrt{\widehat{\text{MSE}}}$ intervals at a fine grid of points; and the dashed-line curve is the true surface. Because this is stochastic kriging, as opposed to ordinary kriging, the fitted surface need not pass through the data points (see especially at $x = 0.9$), and the $\pm\sqrt{\widehat{\text{MSE}}}$ intervals account both for intrinsic and extrinsic uncertainty about the surface. Notice that the true surface is within the $\pm\sqrt{\widehat{\text{MSE}}}$ bounds on the fitted surface.

The fitted variance curve $\widehat{\mathsf{V}}(x_0)$ is shown in Figure 3. Because we use ordinary kriging for this model, the fitted curve passes through the data points, and it is clear that the simulation provided a particularly poor estimate of $\mathsf{V}(0.9)$.

Using the results from the first-stage experiment (in particular $\widehat{\theta}_\mathsf{M}$ and $\widehat{\mathsf{V}}(x)$), we apply (29) to obtain the optimal allocation of $N = 500$ replications to the full set of design points $x = 0.3$, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The variance model is required because the full design includes design points that were not simulated in the first-stage experiment. The estimated optimal allocation is $n = 2$, 80, 11, 81, 33, 165, 128, respectively. That design points 2 and 4 (0.4 and 0.6) receive relatively large allocations relative to design points 1, 3, and 5 (0.3, 0.5, and 0.7) results mostly from their variance being overestimated by $\widehat{\mathsf{V}}$. More interesting is that $x = 0.8$ receives a larger allocation than $x = 0.9$, even though the standard deviation at 0.9 is predicted to be

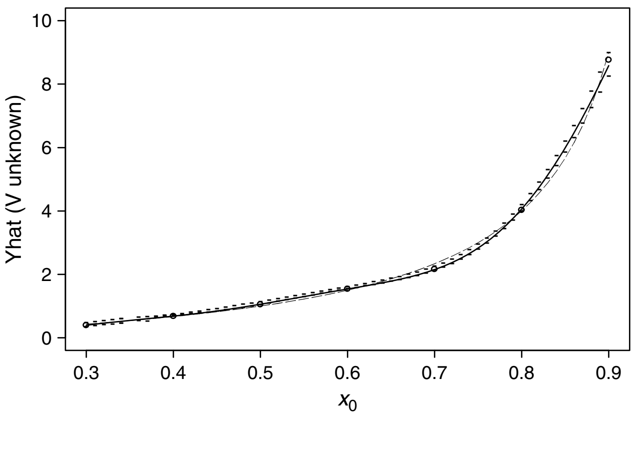
Figure 3. Fitted via ordinary kriging (solid line) and true (dashed line) variance of average number in an M/M/1 queue from the first-stage experiment.



substantially greater than at 0.8 by $\widehat{\mathsf{V}}$. This occurs because our optimal allocation considers not only the relative standard deviations at the design points, but also their range of influence in the metamodel; $x = 0.8$ is closer to more points in the design than 0.9, and therefore is more valuable.

Because several of the design points have already received more replications than the optimal allocation above—always a danger when the initial sample size has to be selected arbitrarily—we ran the second-stage experiment allocating the 500 replications optimally (in practice we would not discard the data we already have and would instead allocate as close to the optimal design as possible). Figure 4 shows the result. The most important thing to notice is not the close fit to the true curve as much as the nearly constant $\pm\sqrt{\widehat{\mathrm{MSE}}}$ intervals surrounding the fitted curve.

Figure 4. Fitted via stochastic kriging (solid line) and true (dashed line) expected number in an M/M/1 queue from the second-stage experiment.



To assess the penalty for estimating the intrinsic variance, and also our ability to capture the error in our response surface, we made M macroreplications of the entire procedure, applying both the full stochastic kriging estimator and the stochastic kriging estimator using the known function $\mathsf{V}(x)$, and computed

$$\overline{\mathrm{IMSE}} = \frac{1}{M}\sum_{l=1}^{M}\left\{\int_{x_L}^{x_U}\left(\widehat{\widehat{\mathsf{Y}}}_l(x_0) - \frac{x}{1-x}\right)^2 dx_0\right\}$$

and

$$\widehat{\mathrm{IMSE}} = \frac{1}{M}\sum_{l=1}^{M}\left\{\int_{x_L}^{x_U}\widehat{\mathrm{MSE}}_l(x_0)\,dx_0\right\},$$

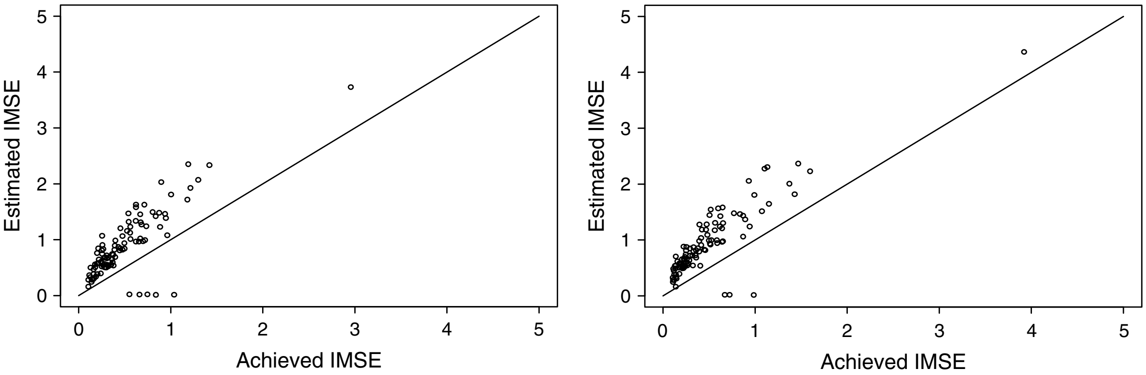
where the subscript l denotes the fit from the lth macroreplication. The quantity $\overline{\mathrm{IMSE}}$ is an unbiased estimator of the achieved integrated MSE; we compare the $\overline{\mathrm{IMSE}}$ with and without using the known variance V to assess the impact of estimating it. The quantity $\int_{x_L}^{x_U}\widehat{\mathrm{MSE}}_l(x_0)\,dx_0$ is an internal estimator of the integrated MSE; we compare $\widehat{\mathrm{IMSE}}$ with $\overline{\mathrm{IMSE}}$ to evaluate how well our internal estimator of MSE performs, and also look at the individual values graphically.

We obtained these comparisons for a number of different experiment designs and observed the following: Our internal estimator $\widehat{\mathrm{MSE}}$ tends to overestimate the true MSE, but occasionally substantially underestimates it, usually when the number of design points or number of replications at the design points is quite small. There appeared to be little or no penalty for estimating V, and in many cases the achieved IMSE was actually smaller when we estimated V rather than using the known function.

To take one representative example, consider the first-stage design described earlier in this section: Design points $x = 0.3, 0.5, 0.7, 0.9$, with $n = 20$ replications at each point of length $T = 1,000$ time units. We made $M = 100$ macroreplications of the experiment and estimated the IMSE over the range $x_L = 0.3$ to $x_U = 0.9$. The achieved $\overline{\mathrm{IMSE}}$ was 0.508 versus 0.503 (with standard error 0.04) using estimated versus known V, respectively. The corresponding $\widehat{\mathrm{IMSE}}$ values were 0.893 and 0.943, showing that we overestimated IMSE on average. Figure 5 is a scatterplot of $\int_{x_L}^{x_U}(\widehat{\widehat{\mathsf{Y}}}_l(x_0) - x/(1-x))^2\,dx_0$ versus $\int_{x_L}^{x_U}\widehat{\mathrm{MSE}}_l(x_0)\,dx_0$ for all 100 macroreplications and the cases of unknown and known V. The general trend of overestimation is clear, with a few trials in which there was substantial underestimation. Adding design points improves performance, but clearly additional work on MSE estimation is required.

Of course, the M/M/1 queue is just one illustration, so general conclusions cannot be reached other than a strong suggestion that stochastic kriging is behaving as theory suggests.

Figure 5. Scatterplot of achieved vs. estimated MSE when variance is estimated (left) or known (right).



5. Conclusions

This paper provides a mathematical foundation for stochastic kriging, a method that extends the power of kriging metamodeling for deterministic computer experiments to modeling responses from stochastic simulations. To realize the full potential of this technique, we need to, and are, addressing these follow-up issues:

1. Our initial results on experiment design should lead to methods for sequential, adaptive design that places design points and allocates simulation effort as we learn more about the response surface being modeled. The ability to capture intrinsic and extrinsic uncertainty in the design is a strength of stochastic kriging.

2. In our limited experiments it appeared that the Gaussian random field model with Gaussian correlation structure did not work as well for representing estimator variance as it did for the response mean. Other alternative models should be explored, as well as whether there is any benefit from fitting a joint model for (M, V).

3. We largely ignored the possibility of including a trend term, $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$, in our metamodel. Clearly there are applications for which the form of such a term is known or suspected, and including it may leads to better fits. The presence of a trend term may make the use of CRN worthwhile.

4. The examples in this paper employed only a one-dimensional design variable x, but the theory is for general d-dimensional $\mathbf{x}$. In addition to the numerical issues that can arise in fitting high-dimensional kriging models, there is also a practical matter of visualizing and exploring the fitted surface. Tools such as ATSV (Stump et al. 2007) may be particularly helpful in this regard.

6. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnote

1. This example suggests how we might apply stochastic kriging in a single-run, steady-state simulation, but we do not go any deeper into that topic here. The computational cost of simulating t units of time depends on the structure of the simulation algorithm, making it difficult to provide general results. For example, here we treat the computational cost of simulating t units of time as t, but it could also depend on x.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under grant DMI-0555485, by the Semiconductor Research Corporation under grant 2004-OJ-1225, and by General Motors R&D. The authors also acknowledge helpful advice from Dan Apley, Russell Barton, Thomas Santner, and Tim Simpson. They thank the associate editor, two anonymous referees, and Evren Baysal for helpful comments and corrections. Portions of this paper were published in Ankenman et al. (2008).

References

Ankenman, B., B. L. Nelson, J. Staum. 2008. Stochastic kriging for simulation metamodeling. *Proc. Winter Simulation Conf.* IEEE, Piscataway, NJ, 362–370.

Barton, R. R., M. Meckesheimer. 2006. Metamodel-based simulation optimization. S. G. Henderson, B. L. Nelson, eds. *Elsevier Handbooks in Operations Research and Management Science: Simulation*, Chapter 19. Elsevier, New York.

Biles, W. E., J. P. C. Kleijnen, W. C. M. van Beers, I. Nieuwenhuyse. 2007. Kriging metamodeling in constrained simulation optimization: An exploratory study. *Proc. 2007 Winter Simulation Conf.* IEEE, Piscataway, NJ, 355–362.

Boesel, J., B. L. Nelson, S. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. *Oper. Res.* **51** 814–825.

Cheng, R. C. H., J. P. C. Kleijnen. 1998. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Oper. Res.* **47** 762–777.

Cressie, N. A. C. 1993. *Statistics for Spatial Data*. Wiley, New York.

Fang, K. T., R. Li, A. Sudjianto. 2006. *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, Boca Raton, FL.

Kennedy, M. C., A. O'Hagan. 2000. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87** 1–13.

Kleijnen, J. P. C., W. C. M. van Beers. 2005. Robustness of Kriging when interpolating in random simulation with heterogeneous variances: Some experiments. *Eur. J. Oper. Res.* **165** 826–834.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, New York.

Mitchell, T. J., M. D. Morris. 1992. The spatial correlation function approach to response surface estimation. *Proc. 1992 Winter Simulation Conf.* IEEE, Piscataway, NJ, 565–571.

Myers, R. H., D. C. Montgomery. 2002. *Response Surface Methodology*, 2nd ed. Wiley, New York.

Nelson, B. L. 1990. Control-variate remedies. *Oper. Res.* **38** 974–992.

Nozari, A., S. F. Arnold, C. D. Pegden. 1987. Statistical analysis for use with the Schruben and Margolin correlation induction strategy. *Oper. Res.* **35** 127–139.

O'Hagan, A., J. J. Forster. 2004. *Bayesian Inference*, 2nd ed. *Kendall's Advanced Theory of Statistics, Volume 2B*. Oxford University Press, London.

Sabuncuoglu, I., S. Touhami. 2002. Simulation metamodeling with neural networks: An experimental investigation. *Internat. J. Production Res.* **40** 2483–2505.

Sacks, J., W. J. Welch, T. J. Mitchell, H. P. Wynn. 1989. Design and analysis of computer experiments. *Statist. Sci.* **4** 409–423.

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

Schruben, L. W., B. H. Margolin. 1978. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *J. Amer. Statist. Assoc.* **73** 504–525.

Searle, S. R., G. Casella, C. E. McCulloch. 1992. *Variance Components*. Wiley, New York.

Stein, M. L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Stump, G., S. Lego, M. Yukish, T. W. Simpson, J. A. Donndelinger. 2007. Visual steering commands for trade space exploration: User-guided sampling with example. F. Liou, ed. *ASME Design Engineering Technical Conferences—Design Automation Conf.* ASME, Las Vegas, NV, September 4–7, DETC2007/DAC-34684.

Tew, J. D., J. R. Wilson. 1992. Validation of simulation analysis methods for the Schruben-Margolin correlation-induction strategy. *Oper. Res.* **40** 87–103.

Tew, J. D., J. R. Wilson. 1994. Estimating simulation metamodels using combined correlation-based variance reduction techniques. *IIE Trans.* **26** 2–16.

van Beers, W. C. M., J. P. C. Kleijnen. 2003. Kriging for interpolation in random simulation. *J. Oper. Res. Soc.* **54** 255–262.

van Beers, W. C. M., J. P. C. Kleijnen. 2008. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *Eur. J. Oper. Res.* **186** 1099–1113.

Whitt, W. 1989. Planning queueing simulations. *Management Sci.* **35** 1341–1366.

Yang, F., B. E. Ankenman, B. L. Nelson. 2007. Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Res. Logist.* **54** 78–93.