*Proceedings of the 2009 Winter Simulation Conference*
*M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.*

# BETTER SIMULATION METAMODELING: THE WHY, WHAT, AND HOW OF STOCHASTIC KRIGING

Jeremy Staum

Department of Industrial Engineering and Management Sciences
McCormick School of Engineering
Northwestern University
2145 Sheridan Road
Evanston, IL 60208-3119, U.S.A.

## ABSTRACT

Stochastic kriging is a methodology recently developed for metamodeling stochastic simulation. Stochastic kriging can partake of the behavior of kriging and of generalized least squares regression. This advanced tutorial explains regression, kriging, and stochastic kriging as metamodeling methodologies, emphasizing the consequences of misspecified models for global metamodeling. It provides an exposition of how to choose parameters in stochastic kriging and how to build a metamodel with it given simulation output, and discusses future research directions to enhance stochastic kriging.

## 1 INTRODUCTION: SIMULATION METAMODELING

By running a stochastic simulation, we can learn about a quantity (e.g. the expected waiting time of a customer in a queueing system) that is specified by a stochastic model, but which we can not compute analytically. Often, we are interested in learning about how this quantity varies as a function $y$, called the *response surface*, of some inputs $x$ to the simulation model. In this tutorial, we work with the following example.

**Example 1** *We simulate an M/M/1 queue with arrival rate 1 and service rate x. The steady-state waiting time is positive with probability* $1/x$ *and, given that it is positive, is conditionally exponential with mean* $1/(x-1)$*. Its mean is* $y(x) = 1/(x(x-1))$*. Each simulation run is initialized in steady state (which avoids bias from the initial conditions) and simulates a fixed number of customers. Its output is their average waiting time.*

This model is simple, but it illustrates some of the key features that are commonly found in simulation models used in operations research but which are non-standard in some relevant fields of statistics:

- The response surface $y$ is smooth and monotone.
- The variability of the response surface is much greater over some parts of the domain than others: the steady-state expected waiting time varies from 9.09 to 4.17 for $x \in [1.1, 1.2]$ but it varies from 0.585 to 0.5 for $x \in [1.9, 2]$.
- The variance of the simulation output is much greater over some parts of the domain than others: it is two orders of magnitude larger for $x$ near 1.1 than for $x$ near 2.

A stochastic simulation run at the input $x$ with effort $n$ provides an estimate $Y(x;n)$ of $y(x)$. The effort $n$ is related to the number of independent replications or to run length in steady-state simulation. In this tutorial, we focus on *global metamodeling*, approximating the response surface $y$ over a large domain $\mathscr{X}$. For the sake of providing decision support, we might want to be able to provide an accurate estimate of $y(x)$ for any *prediction point* $x \in \mathscr{X}$, as soon as a decision-maker asks about $x$, without waiting to run a simulation with sufficiently large effort $n$ at $x$. Or we might want to know the value of $y$ at many prediction points, but it would be too slow to run lengthy simulations at all of them to estimate separately the value of the response surface at each prediction point.

1. **Experiment design:** Choose *design points* $x_1, \ldots, x_k$ and the associated simulation effort $n_1, \ldots, n_k$.
2. **Simulation:** For $i = 1, \ldots, k$, perform a simulation with effort $n_i$ at design point $x_i$.
3. **Metamodeling:** Using the simulation outputs $Y(x_1), \ldots, Y(x_k)$, build the metamodel $\widehat{y}$.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

We treat the metamodel $\widehat{y}$ as an approximation of the response surface $y$: for any prediction point $x \in \mathscr{X}$ in which we are interested, we *predict* $y(x)$ by $\widehat{y}(x)$. This can be viewed as a prediction of the simulation output $Y(x;n)$ we would get if we ran the simulation model at $x$ with a large effort $n$.

In this tutorial, we focus on Step 3, exploring alternative methodologies for global metamodeling: regression, kriging, and stochastic kriging, recently proposed by Ankenman, Nelson, and Staum (2008). In operations research, because response surfaces are commonly smooth almost everywhere, but nonlinear with an unknown functional form, it is much easier to do good local metamodeling than global metamodeling. Local metamodeling based on regression has been successful in optimization via stochastic simulation (Barton and Meckesheimer 2006, Kleijnen 2008). For global metamodeling, kriging has been more successful, as reported by Kleijnen (2008, 2009). The purpose of stochastic kriging is to improve upon kriging as a global metamodeling methodology for stochastic simulation. Because the aim of this tutorial is to give an exposition of stochastic kriging, and little is known about experiment design for stochastic kriging, we merely touch on this important topic in Section 7.2.

## 2 RANDOM FIELDS

How is it possible to infer anything about $y(x)$ if no simulation has been performed at $x$? Some assumptions about the simulation model must be made to justify a metamodeling procedure, especially assumptions about the response surface $y$, such as continuity or differentiability. There is a literature on function approximation when $y$ can be evaluated exactly, as in deterministic simulations, and a literature on filtering when the observations of $y$ are contaminated with noise, as happens in stochastic simulations. One way to view the situation, appropriate for stochastic simulations, is in terms of random fields, which are essential to kriging.

A *random field* is a function $\mathscr{M} : \mathscr{X} \times \Omega \to \mathbb{R}$, where $\Omega$ is a sample space that has a probability measure on it, as usual in probability theory. For each $\omega \in \Omega$, the realization $\mathscr{M}_\omega$ is a function from $\mathscr{X}$ to $\mathbb{R}$. For each $x \in \mathscr{X}$, $\mathscr{M}(x)$ is a random variable. A familiar class of random fields has $\mathscr{X} = \mathbb{R}_+$ and $x \in \mathbb{R}_+$ is given the interpretation of time, so that $\mathscr{M}(x)$ is the value observed at time $x$: these random fields are stochastic processes, e.g. Brownian motion and continuous-time Markov chains, and their realizations are called “sample paths.” The *mean function* of the random field $\mathscr{M}$ maps an input $x$ to the expected value $\mathrm{E}[\mathscr{M}(x)]$ of the random field observed at $x$. Likewise, the *covariance function* of the random field maps a pair of points $(x,x')$ to $\mathrm{Cov}[\mathscr{M}(x),\mathscr{M}(x')]$.

One random field pertinent to simulation modeling is $\mathscr{Y}$ where $\mathscr{Y}(x)$ is the output of a single simulation replication run at $x$ and common random numbers (CRN) are used in simulations run at different points. If the simulation is unbiased, the mean function of $\mathscr{Y}$ is $y$. The covariance function of $\mathscr{Y}$ is $c$ where $c(x,x') = \rho(x,x')\sqrt{v(x)v(x')}$ and $\rho(x,x')$ is the correlation induced by CRN between simulation outputs at $x$ and $x'$. The output of a simulation that uses independent random numbers at all points is a different random field, $\mathscr{Y}_\perp$. The random fields $\mathscr{Y}$ and $\mathscr{Y}_\perp$ have the same marginal distributions, hence the same mean function, but different joint distributions and different covariance functions. The covariance function of $\mathscr{Y}_\perp$ maps $(x,x')$ to 0 if $x \neq x'$, and maps $(x,x)$ to $v(x) = \mathrm{Var}[\mathscr{Y}(x)]$. Some realizations of these two random fields appear in the left ($\mathscr{Y}$) and right ($\mathscr{Y}_\perp$) panels of Figure 1, based on Example 1 with a run length of 1. That is, $\mathscr{Y}(x)$ is a single draw from the steady-state distribution of waiting time when the service rate is $x$. This can be accomplished by sampling $U_1$ and $U_2$ independently and uniformly on $(0,1]$, and letting $\mathscr{Y}(x) = -(x-1)^{-1}(\ln U_1)1\{U_2 < 1/x\}$. Similarly, $\mathscr{Y}_\perp(x) = -(x-1)^{-1}(\ln U_1(x))1\{U_2(x) < 1/x\}$, where $U(x)$ is independent of $U(x')$ whenever $x \neq x'$. The realizations of $\mathscr{Y}_\perp$ are plotted only at 10 distinct points so that the plot is legible; $\mathscr{Y}_\perp$ is almost surely discontinuous everywhere, whereas $\mathscr{Y}$ is discontinuous at one point at most, differentiable everywhere else, and nonincreasing.

The random field $\mathscr{Y}$ has been discussed in the simulation literature, although it is not always called a random field. For example, $\mathscr{Y}$ is studied in optimization via simulation to provide conditions ensuring the continuity of $y$ (Kim and Henderson 2008). Differentiating $\mathscr{Y}$ with respect to $x$ is the basis for the infinitesimal perturbation analysis (IPA) method of estimating derivatives of $y$ (Fu 2008). Borogovac and Vakili (2008) show that when $\rho(x,x')$ is large for $x'$ near $x$, it is possible to achieve large variance reduction in estimating the value of the response surface $y$ at many nearby points; they also consider using derivatives of $\mathscr{Y}$ in their method. If a simulation at $x$ provides estimates of the response $y(x)$ and of its derivatives, this can improve metamodeling (Morris, Mitchell, and Ylvisaker 1993).

We will be particularly concerned with *Gaussian random fields* (GRFs). A random field $\mathscr{M}$ is a GRF if its finite-dimensional distributions are Gaussian: that is, for any finite set $x_1,\dots,x_K$, $[\mathscr{M}(x_1),\dots,\mathscr{M}(x_K)]$ is $K$-variate normal. A $K$-variate normal random vector with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$ is a GRF on $\{1,\dots,K\}$ with mean function $\mu$ given by $\mu(i) = m_i$ and covariance function $\sigma^2$ given by $\sigma^2(i,j) = \Sigma_{ij}$. Brownian motion is a GRF on $\mathbb{R}_+$ with mean 0 and covariance function $\sigma^2$ given by $\sigma^2(x,x') = \min\{x,x'\}$.

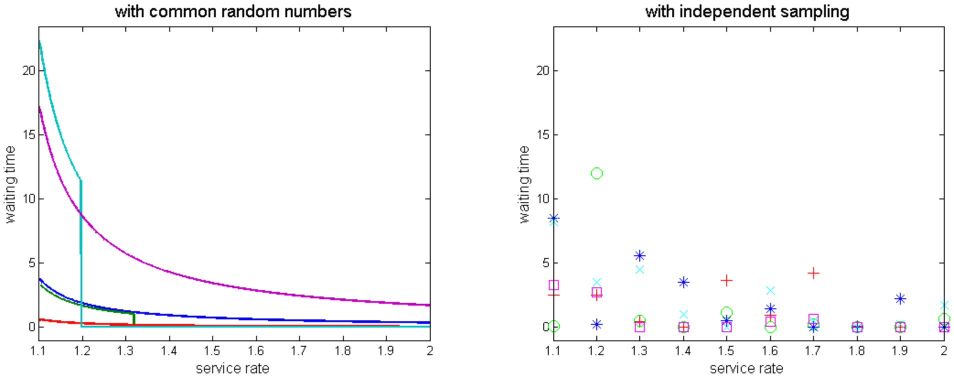

Figure 1: Five realizations of the random fields $\mathscr{Y}$ and $\mathscr{Y}_\perp$ representing output of the M/M/1 simulation (Example 1).

## 3 REGRESSION

An assumption commonly used to justify and analyze ordinary least-squares (OLS) regression is

$$Y(x) = y(x) + \varepsilon(x), \quad y(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \varepsilon \sim \mathrm{WN}(v) \tag{1}$$

where $\mathrm{WN}(v)$ is white noise with variance $v$, the GRF with mean function 0 and covariance function $c$ given by $c(x,x) = v$ and $c(x,x') = 0$ for $x \neq x'$: the errors at different points are assumed to be independent. Equivalently, $Y \sim \mathrm{GRF}(y,c)$. The other assumption in (1) is that the response surface $y$ equals a linear combination of the components of a known function $\boldsymbol{b} : \mathbb{R}^d \to \mathbb{R}^q$. We call the function $\boldsymbol{b}(\cdot)\boldsymbol{\beta} : \mathbb{R}^d \to \mathbb{R}$ the *trend*. In regression $\boldsymbol{\beta}$ is regarded as unknown and something to be estimated. The OLS estimator is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{B}^\top \boldsymbol{B})^{-1}(\boldsymbol{B}^\top \boldsymbol{Y})$ where the row $\boldsymbol{B}_{i\cdot} = \boldsymbol{b}(x_i)$ and $\boldsymbol{Y} = [Y(x_1), \ldots, Y(x_k)]^\top$. The metamodel is the estimated trend: $\widehat{y}(x) = \boldsymbol{b}(x)\widehat{\boldsymbol{\beta}}$ for any $x \in \mathscr{X}$.

The OLS estimator $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator (MLE): the value of $\boldsymbol{\beta}$ that maximizes the log-likelihood $-\left(k\ln(2\pi) + (\ln v) + \|\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}\|^2/v\right)/2$ of the data according to (1) is the value that minimizes the sum of squared residuals $\|\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}\|^2 = \sum_{i=1}^k (Y(x_i) - \boldsymbol{b}(x_i)\boldsymbol{\beta})^2$. Remarkably, this value $\widehat{\boldsymbol{\beta}}$, and hence the metamodel, do not depend on the variance $v$, or even on whether the variance is known. This point will be important in our discussion of weighted least squares. If the assumptions in (1) held, the MLE and metamodel would have good statistical properties, as explained in textbooks on regression such as Weisberg (1985).

*Model misspecification* is the failure of the assumptions to describe the data well. It can cause extremely poor prediction. The statistics literature discusses model misspecification amply, including diagnostics and remedies. We emphasize model misspecification to frame and illustrate the advantages and disadvantages of stochastic kriging. In global metamodeling, we are seldom readily able to find a function $\boldsymbol{b}$ such that $y(x)$ is nearly $\boldsymbol{b}(x)\boldsymbol{\beta}$ for all $x \in \mathscr{X}$ and some $\boldsymbol{\beta}$. In simulation metamodeling, heterogeneous variances of simulation outputs, and correlations induced between them by common random numbers, can violate the white-noise assumption. Barton and Meckesheimer (2006) and Kleijnen (2008) discuss remedies for this in the context of simulation, principally generalized least squares and variance-stabilizing transformations.

Generalized least squares (GLS) drops the white-noise assumption and allows dependence among the errors (Weisberg 1985, § 6.2). This amounts to replacing $\varepsilon \sim \mathrm{WN}(v)$ in (1) with $\varepsilon \sim \mathrm{GRF}(0,c)$, where the more general covariance function $c$ determines the covariance matrix $\boldsymbol{C}$ of the data $\boldsymbol{Y}$. The log-likelihood of the data is then

$$-\frac{1}{2}\left(k\ln(2\pi) + \ln(|\boldsymbol{C}|) + (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta})^\top \boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta})\right)$$

and the MLE of $\boldsymbol{\beta}$ is the GLS estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{B}^\top \boldsymbol{C}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^\top \boldsymbol{C}^{-1}\boldsymbol{Y}. \tag{2}$$

The estimator, and hence the metamodel $\widehat{y}(\cdot) = \boldsymbol{b}(\cdot)\widehat{\boldsymbol{\beta}}$, do depend on the covariance matrix $\boldsymbol{C}$—but not on its magnitude! That is, the GLS estimator (2) is invariant to multiplying $\boldsymbol{C}$ by a positive number. This explains the workings of weighted least squares (WLS), a special case of GLS in which the data is assumed to be independent, so that $c(x,x') = 0$ for $x \neq x'$ and $\boldsymbol{C}$ is diagonal. The WLS estimator is the same whether we use $\boldsymbol{C}$ or divide it by its trace, which amounts to assigning to each data point a weight inversely proportional to its variance, such that the weights sum to one. Only the relative magnitudes of the variances matter.

In simulation metamodeling, we can estimate the covariances of the simulation outputs. Suppose there are multiple replications, and let $Y_j(x)$ be the output of replication $j$ of a simulation at $x$. As discussed in Section 2, $\{Y_j(\cdot)\}_{j \in \mathbb{N}}$ is a sequence of independent realizations of the random field $\mathscr{Y}$ (with CRN) or $\mathscr{Y}_\perp$ (with independent sampling). We focus on independent sampling, the case of WLS. Then we can replace the assumption (1) with the assumption

$$Y_j(x) = y(x) + \varepsilon_j(x), \quad y(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \mathrm{E}[\varepsilon_j(x)] = 0, \quad \mathrm{Var}[\varepsilon_j(x)] = v(x_j), \quad \varepsilon_j(x_h) \perp \varepsilon_{j'}(x_i) \text{ if } j \neq j' \text{ or } h \neq i. \tag{3}$$

If $n_1, \ldots, n_k$ are large enough to invoke the central limit theorem, this supports the conclusion that

$$\boldsymbol{Y} \sim \mathscr{N}(\boldsymbol{B\beta}, \boldsymbol{C}), \quad \boldsymbol{C}_{ii} = v(x_i), \quad \boldsymbol{C}_{hi} = 0 \text{ if } h \neq i. \tag{4}$$

Typically, we do not know the variances of the simulation output, so we estimate them: let $\mathscr{S}^2(x_i)$ be the sample variance of $Y_1(x_i), \ldots, Y_{n_i}(x_i)$. Empirical weighted least squares (EWLS) assigns to design point $x_i$ a weight inversely proportional to the estimate $\mathscr{S}^2(x_i)/n_i$ of $\boldsymbol{C}_{ii}$.

Figure 2 shows the output $\boldsymbol{Y}$ of a simulation of Example 1, as well as metamodels built by OLS and EWLS. The top row depicts the output of an experiment with a sparse and deep design, with $k = 6$ design points, each having 30 replications, each of which is a simulation run of 1,000 customers. The bottom row depicts the output of an experiment with a dense and shallow design: $k = 60$ design points, each having 30 replications, each of which is a simulation run of 100 customers. The left and center columns show the predictions of misspecified metamodels that assume the expected waiting time $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, quadratic in the service rate. The metamodels' predictions are poor both in absolute terms (left column, showing $\widehat{y}$) and relative to the true values (center column, showing $\widehat{y}/y - 1$). This is not because $\boldsymbol{\beta}$ is estimated poorly, but because there is no function quadratic in the service rate that approximates the response surface $y$ well. The quadratic function $f$ that best approximates $y$ in the least squares sense over this domain, i.e. minimizes $\int_{1.1}^{2} (f(x) - y(x))^2 \, dx$, is plotted as a dotted line. The choice $\boldsymbol{b}(x) = [1, x, x^2]$ does not support accurate prediction.

The failure of the misspecified metamodel prompts us to look for a well-specified metamodeling framework. Equation (3) is true as long as we choose $\boldsymbol{b}$ such that there exists $\boldsymbol{\beta}$ satisfying $\boldsymbol{b}(x)\boldsymbol{\beta} = 1/(x(x-1)) = y(x)$. In global metamodeling by regression, we must hope to find a framework such that $\boldsymbol{b\beta} - y$ is small for some $\boldsymbol{\beta}$. It may be easier to think of $\boldsymbol{b}$ such that some transformation of $y$ is nearly a linear combination of its components than to think of $\boldsymbol{b}$ such that $y$ itself is nearly a linear combination of its components. For example, the function $\ln y(x) = -\ln x - \ln(x-1)$ is indeed a linear combination of the functions 1, $\ln x$ and $\ln(x-1)$. The metamodeling framework

$$Y(x) = y(x) + \varepsilon(x), \quad y(x) = \exp\left(\beta_0 + \beta_1 \ln x + \beta_2 \ln(x-1)\right), \quad \varepsilon \sim \mathrm{GRF}(0, c) \tag{5}$$

is well-specified, again supposing that the simulation effort is large enough that the errors are indeed approximately normal. Equation (5) is a GLM, a *generalized linear model* (McCullagh and Nelder 1989): it represents the response surface $y$ as a known nonlinear function of a linear combination with unknown coefficients $\boldsymbol{\beta}$ of known nonlinear functions of the input $x$, and the errors as being a random field with a known form, but with some unknown parameters (here, their covariances). Because Equation (5) says that $\ln y(x)$ is linear in the unknown coefficients, log is called the *link function* in this GLM. The right column of Figure 2 shows that this well-specified GLM provides metamodels that make much better predictions than those that come from a quadratic metamodel produced by regression.

Figure 2 indicates that, surprisingly, WLS does not necessarily result in better prediction than OLS. When using the well-specified GLM, EWLS provided better predictions by causing better estimation of the regression coefficients $\boldsymbol{\beta}$, in accordance with statistical theory. EWLS was able to do this even though the variances could not be estimated very accurately in the dense and shallow experiment design. The advantages and disadvantages of the EWLS predictions evident in the left column of Figure 2 are inherent to using WLS with a misspecified model. The WLS predictions are good for high service rates and have qualitatively correct behavior: they are positive and monotone, whereas the OLS predictions are non-monotone and, for the sparse and deep experiment design, even negative for some service rates. However, WLS achieves these advantages by giving up on the task of fitting the higher-variance simulation output in the left part of $\mathscr{X}$, where service rates are low. Because the weights in WLS depend only on the ratio of variances, WLS gives up on fitting low service rates even for the sparse and deep experiment design, for which the simulation output variances were all low.

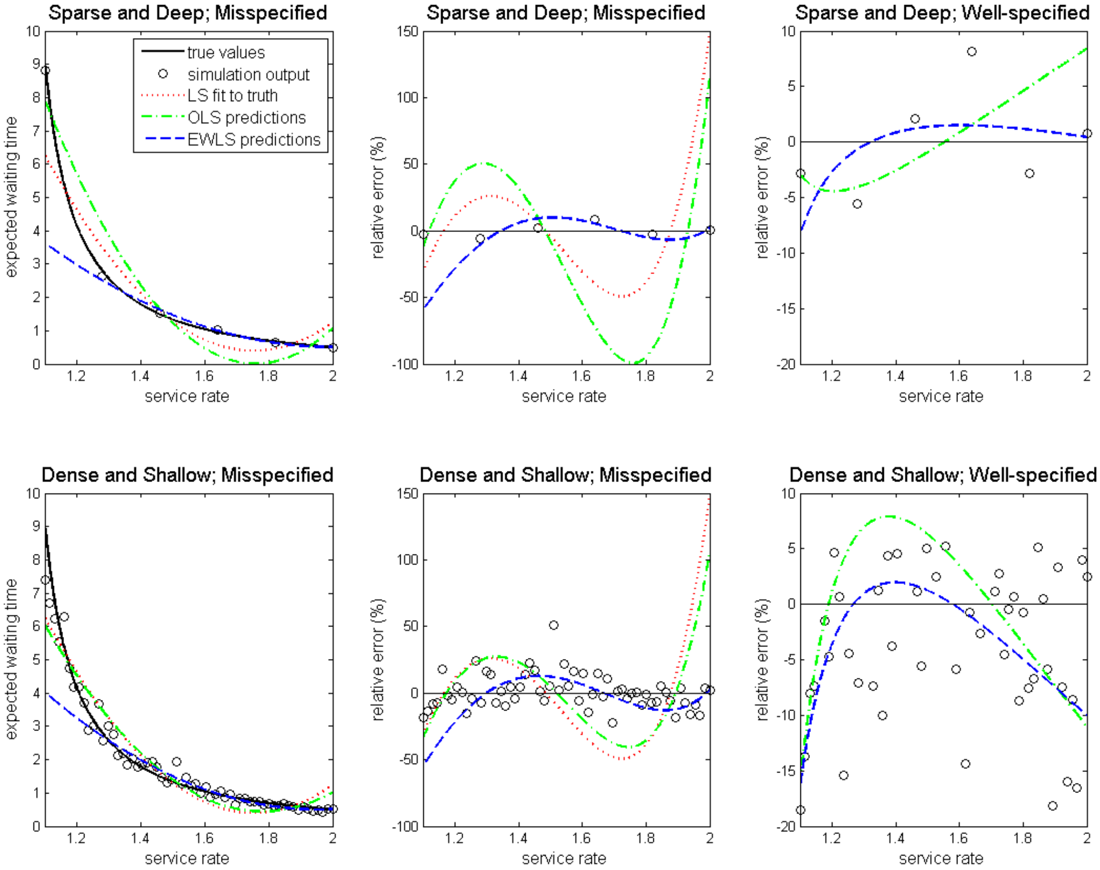

Figure 2: Metamodels of the M/M/1 simulation (Example 1). The rows depict the results of two simulation experiments with different designs. The left and center columns show the predictions of regression based on the assumption that expected waiting time is quadratic in service rate. The right column uses the well-specified generalized linear model in Equation (5).

As a simple illustration of this general point, consider fitting a constant metamodel, $\widehat{y}(x) = \beta_0$, to the response surface $y(x) = x$ on the domain $\mathscr{X} = [0,1]$, using simulation outputs $Y(0) = 0$ and $Y(1) = 1$ with variances $v(1) = 4v(0)$. However small $v(1)$ is, the output $Y(1) = 1$ receives weight 0.2, and the WLS prediction is $\widehat{\beta}_0 = 0.2$, which is worse in the least-squares sense than the OLS prediction of $\widehat{\beta}_0 = 0.5$. If the simulation output variances are small, WLS is inappropriate because it gives little weight to $Y(1)$, which contains high-quality information.

Variance-stabilizing transformations (VSTs) transform the simulation output to make the variance of the error more nearly constant across all $x \in \mathscr{X}$ (Weisberg 1985, § 6.2). For example, we might modify Equation (1) to

$$\ln Y(x) = \ln y(x) + \varepsilon'(x), \quad \ln y(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \varepsilon' \sim \mathrm{WN}(v). \tag{6}$$

Log transformation of the data is appropriate when the variance of the error $\varepsilon(x) = Y(x) - y(x)$ is proportional to the square of the response $y(x)$, and indeed the steady-state variance of the waiting time is equal to the square of the steady-state mean waiting time $y(x)$ in Example 1. This VST is very closely related to a GLM: Equation (5) also lets $\ln y$ be a linear combination, with unknown coefficients, of known functions. The difference is in the treatment of errors. The GLM correctly models the untransformed errors $\varepsilon(x) = Y(x) - y(x)$ as having zero mean; moreover, they are approximately normal for sufficiently large simulation effort. Thus $\ln Y(x) - \ln y(x) = \ln(1 + \varepsilon(x)/y(x))$ has negative mean and hence Equation (6) is incorrect in saying $\varepsilon'(x)$ has zero mean. This can cause bias in prediction, as shown in the right panel in Figure 3, which compares the use of a VST to the similar GLM, using OLS and EWLS. Whereas the weight of $Y(x_i)$ in EWLS with the

GLM is inversely proportional to the sample variance of the $n_i$ replications run at $x_i$, the weight of $\ln Y(x_i)$ in EWLS with the VST is inversely proportional to a variance estimated by the delta method (Asmussen and Glynn 2007, p. 75).

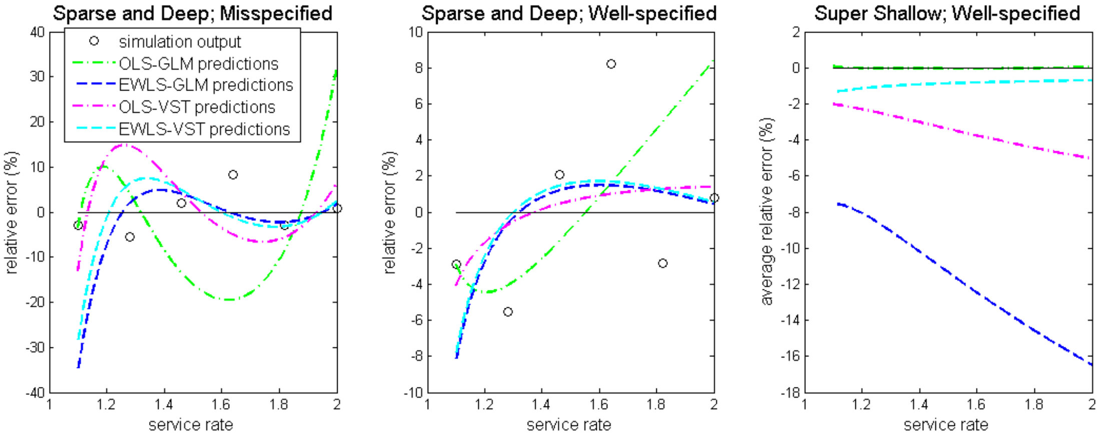

Figure 3: Metamodels of the M/M/1 simulation (Example 1), using log as the link function in a GLM or as a variance-stabilizing transformation. The left panel shows predictions based on the assumption that the log expected waiting time is quadratic in service rate. The center and right plots are based on the well-specified model (5). The right panel shows the average relative error of predictions over 2,000 macro-replications of a simulation experiment with high output variance.

The left and center panels in Figure 3 use the experiment design with $k = 6$ design points, 30 replications each, and 1,000 customers per replication. The left panel features the misspecified model $\beta_0 + \beta_1 x + \beta_2 x^2$ for $\ln y(x)$. The resulting predictions are poor, but much better than those that come from modeling $y(x)$ as quadratic in $x$ (top center panel of Figure 2). The center panel in Figure 3 is the same as the top right panel in Figure 2, but with lines added for the VST as well as the GLM, all providing good predictions based on the well-specified model $\ln y(x) = \beta_0 + \beta_1 \ln x + \beta_2 \ln(x-1)$. In these plots, the VST is minimizing errors on the log scale (equivalent to relative error) whereas the GLM is minimizing errors on the original scale, that is, regarding a deviation of 0.5 from $Y(1.1)$, which is large, as just as bad as a deviation of 0.5 from $Y(2)$, which is small. Although the overall relative error of prediction by OLS-GLM in the left panel is worse than the others, its error in absolute terms is better. (EWLS-GLM performs similarly to EWLS-VST in this example because it assigns low weights to outputs with low service rates, and then fits the other outputs well.) The right panel in Figure 3 uses the same well-specified model and shows the average relative error in 2,000 macro-replications of a simulation experiment of Example 1 with $k = 60$ design points, 30 replications each, and only one customer per replication. Because of initialization in steady state, each simulation output is the average of 30 independent draws from the steady-state distribution of customer waiting time, and this has very large variance. The purpose is to see whether such large variance causes problems with bias for VST and for output variance estimation for EWLS. The answer to both questions is yes, but the estimated bias for VST is small. EWLS generates a substantial bias due to unreliable choice of weights. It is not so bad for VST as for GLM, precisely because the VST stabilizes variance.

As an overview, we can say about regression for global metamodeling of simulation:

- Predictions tend to be inaccurate when the model is misspecified. However, it can take a lot of analyst time to find and validate a well-specified model, especially in high dimension.
- Generalized linear models or variance-stabilizing transformations can help to deal with nonlinear relationships and heterogeneous simulation output variances.
- Weighted least squares is dangerous when applied with misspecified regression models. The safer strategy would be to use an experiment design, perhaps a two-stage experiment design (Kleijnen 2008, § 3.4.5), that makes all the simulation output variances low, and then to use ordinary least squares.
- Experiment designs that result in very high simulation output variances can cause problems for empirical weighted least squares and variance stabilizing transformations.

Cheng and Kleijnen (1999) address such issues in metamodeling of queueing simulations and associated design of experiments.

## 4 KRIGING

The main idea of kriging is to regard the response surface $y$ as a realization of a GRF $\mathsf{Y}$. In its most common form, kriging is based on the assumption

$$Y(x) = y(x) = \mathsf{Y}(x), \quad \mathsf{Y} \sim \mathrm{GRF}(\mu, \sigma^2), \quad \mu(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \sigma^2(x,x') = \tau^2 r(x-x';\boldsymbol{\theta}), \tag{7}$$

where the parameters $\boldsymbol{\beta}$, $\tau^2$, and $\boldsymbol{\theta}$ are to be estimated, while $\boldsymbol{b}$ is known and the *spatial correlation function* $r$ is known up to the parameter $\theta$. A covariance function $\sigma^2$ with the structure in (7) is *stationary*: $\mathrm{Var}[\mathsf{Y}(x)] = \tau^2$ for any $x$, while the correlation between $\mathsf{Y}(x)$ and $\mathsf{Y}(x')$ depends only on the difference $x-x'$. Frequently one takes $\boldsymbol{b} = 1$ so that $\mu(x) = \mathrm{E}[\mathsf{Y}(x)] = \beta_0$ for all $x$, in which case the GRF $\mathsf{Y}$ is stationary because both its mean and covariance functions are stationary. It may seem odd to abandon trend modeling like this; in regression, a good specification of the trend was vital to good prediction.

To contrast kriging with regression, let us suppose that $\boldsymbol{b} = 1$ and the parameters $\boldsymbol{\beta}$, $\tau^2$, and $\boldsymbol{\theta}$ are known. If we assumed, in the context of regression, that $\boldsymbol{b} = 1$ and $\boldsymbol{\beta}$ is known, we would be saying that $y(x) = \beta_0$ for all $x$, regardless of the simulation outputs we observe. That would be ridiculous in any interesting example, yet the analogous assumption in the context of kriging can lead to good predictions if the parameters are well-chosen. The difference arises because kriging treats the response surface as a random field while regression treats the response surface as a deterministic trend and treats the error as a random field: cf. Equations (1) and (7). Regression is like filtering, treating a deviation $Y(x) - \boldsymbol{b}(x)\boldsymbol{\beta}$ as uninformative "noise" and aiming to filter it out to learn about the trend, while kriging regards such deviations as informative. Kriging assumes that the simulation output exactly equals the response surface, which can make sense in deterministic computer experiments or physical experiments with negligible measurement error. The meaning of the random field in kriging is completely different from the meaning of the random field in regression: we say the random field $\mathsf{Y} - \mu \sim \mathrm{GRF}(0,\sigma^2)$ in kriging represents *extrinsic* uncertainty because it corresponds to nothing about the simulation model. Instead, it represents the uncertainty we have about the response surface at a point where we have not yet run a simulation.

Continuing to suppose that the parameters $\boldsymbol{\beta}$, $\tau^2$, and $\boldsymbol{\theta}$ are known (but not that $\boldsymbol{b} = 1$), we next consider the predictions that kriging makes. The kriging prediction is

$$\widehat{y}(x) = \widehat{\mathsf{Y}}(x) = \boldsymbol{b}(x)\boldsymbol{\beta} + \boldsymbol{\sigma^2}(x)\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}) = \boldsymbol{b}(x)\boldsymbol{\beta} + \boldsymbol{r}(x)\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}), \tag{8}$$

where $\boldsymbol{\sigma^2}(x)$ is a row vector whose $i$th element is $\sigma^2(x,x_i)$, $\boldsymbol{\Sigma}_{hi} = \sigma^2(x_h,x_i)$, and $\boldsymbol{B}$ has $i$th row $\boldsymbol{B}_{i\cdot} = \boldsymbol{b}(x_i)$. Given a stationary covariance function, the extrinsic variance $\tau^2$ cancels out: $\boldsymbol{r}(x) = \boldsymbol{\sigma^2}(x)/\tau^2$ and $\boldsymbol{R} = \boldsymbol{\Sigma}/\tau^2$. Somewhat as in WLS estimation, the overall level of variance is irrelevant and all that matters is the relative influences of the simulation output at the several design points on $\mathsf{Y}(x)$, determined by the correlation structure $r$. It can be illuminating to take a Bayesian perspective and regard Equation (7) as a prior belief, before seeing the data. The kriging prediction $\widehat{\mathsf{Y}}(x)$ can be viewed either in a Bayesian sense as the posterior expectation of $\mathsf{Y}(x)$ or in a frequentist sense as a best linear unbiased predictor of $\mathsf{Y}(x)$ (Stein 1999). Kriging can be seen as interpolation, where the correlation structure determines the weights: Equation (8) is equivalent to $\widehat{\mathsf{Y}}(x) - \boldsymbol{b}(x)\boldsymbol{\beta} = \boldsymbol{r}(x)\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta})$, which says that the predicted residual (deviation of the response surface from the trend) at $x$ is a weighted sum of the residuals observed at the design points, where the vector $\boldsymbol{r}(x)\boldsymbol{R}^{-1}$ gives the weights. Typically, the weights sum to less than one, so that the predictions also feature some reversion towards the trend. The combination of interpolation and reversion to the trend is appropriate for predictions in geostatistics, the field in which kriging originated: if nearby there is much more gold than average for this region, there is probably not quite as much gold here, but more than average. If the GRF is independent at different locations, i.e. $\sigma^2(x,x') = 0$ for $x \neq x'$, then residuals at the design points are uninformative about the residual at a point that is not a design point, so the kriging prediction is the same as the regression prediction: $\widehat{y}(x) = \boldsymbol{b}(x)\boldsymbol{\beta}$. In this sense, for purposes of prediction, kriging can be viewed as an extension of regression. However, they are founded on quite different assumptions about the response surface and this can be seen from the fact that, whatever the covariance structure, kriging predicts $\widehat{y}(x_i) = \widehat{\mathsf{Y}}(x_i) = \mathsf{Y}(x_i)$ at a design point whereas regression predicts $\widehat{y}(x_i) = \boldsymbol{b}(x_i)\boldsymbol{\beta}$.

Thus, a major issue in getting good predictions from kriging is choosing a good spatial correlation function and value of its parameter $\boldsymbol{\theta}$. Frequently, $r$ is assumed to have a product form:

$$r(x-x';\boldsymbol{\theta}) = \prod_{h=1}^{d} r_0(|x_h - x'_h|;\theta_h). \tag{9}$$

Usually, $r_0$ is a decreasing function that is 1 at 0 and goes to 0 as its argument goes to infinity, so that $\theta_h$ governs the way that correlation decays with distance measured along dimension $h$. The choice of $r_0$ is an important potential source of model misspecification. Stein (1999) discusses families of correlation functions, the consequences of misspecifying them, and advocates using the Matérn family in spatial statistics. The assumption of product form is also an important potential source of model misspecification (Stein 1999, § 2.11), but it is more appropriate in simulation metamodeling, where the axes have different interpretations (e.g. service rate and arrival rate), than in geostatistics. Nonetheless, this means kriging can give different predictions based on the same simulation experiment if the axes are transformed (e.g. to load on the server and arrival rate). Indeed, transforming the scale of a single axis (e.g. from service rate to its log) can change the predictions.

Although the connection to spatial statistics makes it easy to visualize kriging as applying to $\mathscr{X}$ a subset of $\mathbb{R}^d$, it can also be applied when some variables are discrete quantities (e.g. inventory reorder points) or even categorical (e.g. whether back-ordering takes place or not). Discrete and continuous quantities can be handled in the same way. Qian, Wu, and Wu (2008) provide methods for handling categorical variables.

Once the structure of the GRF is specified by choosing $\boldsymbol{b}$ and $r_0$, we use maximum likelihood estimation to choose the parameters $\boldsymbol{\beta}$, $\tau^2$, and $\boldsymbol{\theta}$. Details are presented in Section 6.2 and our implementation relies on the MATLAB function `fmincon` for nonlinear optimization. Figure 4 compares metamodels created by OLS regression and kriging using the exponential correlation function given by $r_0(t;\boldsymbol{\theta}) = \exp(-\boldsymbol{\theta}t)$ and parameters estimated by maximum likelihood. The three experiments all have 30 replications per design point, and the numbers of design points and of customers per replication are 6 and 1,000 (left panel), 20 and 300 (center), and 60 and 100 (right). We use a trend that is quadratic in service rate for regression and for kriging. We also perform kriging a second time with $\boldsymbol{b} = 1$, in which case it estimates the mean level of the response surface over the domain $\mathscr{X} = [1.1, 2]$, not a trend that varies over $\mathscr{X}$.

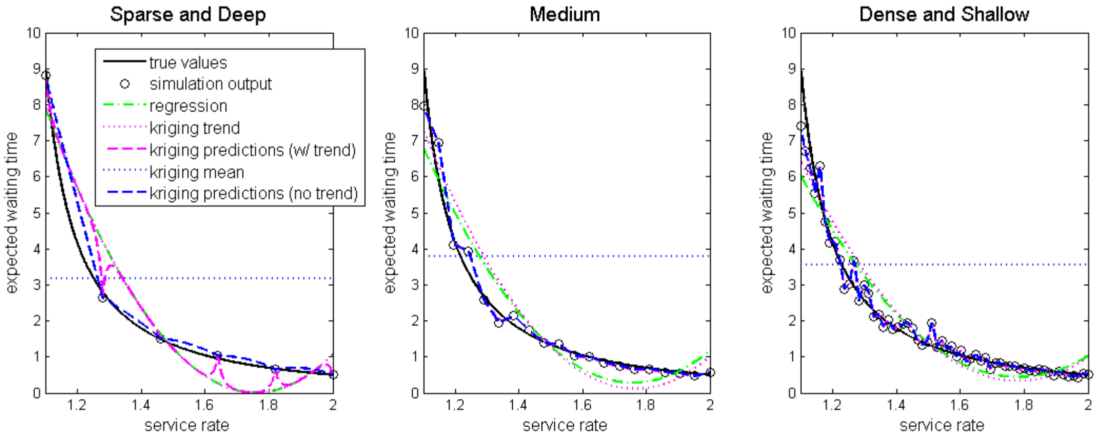

Figure 4: Metamodels of the M/M/1 simulation (Example 1), from simulation experiments with different experiment designs. The trend in expected waiting time is quadratic in service rate.

The maximum likelihood of the simulation output is much greater when a trend model is allowed, but we see from the left panel of Figure 4 that this does not imply that kriging with a trend leads to better predictions! Without trend modeling, it seems likely that there is correlation among the simulation outputs: the leftmost data point is far above the mean, while the other five adjacent data points are below the mean. The MLEs $\widehat{\tau}^2 = 21$ and $\widehat{\boldsymbol{\theta}} = 2.7$ imply there is a lot of variability away from the mean, and a fairly strong degree of correlation: 62% correlation between simulation outputs at adjacent design points. With a quadratic trend, it is not at all clear that there is correlation among the simulation outputs: reading from left to right, one data point is near the trend, then one is below, the next is near the trend, then two are above the trend, and the last one is below. Accordingly, the optimization algorithm terminates with $\widehat{\boldsymbol{\theta}} = 68$, implying negligible correlation, and because the data is much closer to a quadratic trend than to a horizontal line, $\widehat{\tau}^2 = 1.4$. The predictions based on this GRF are much worse: in keeping with our remark above about kriging predictions if simulation outputs are independent, the predictions are nearly the same as those made by regression in most places, but deviating rapidly near design points so that the prediction there equals the simulation output. The experiment design in the left panel has too few design points for quadratic trend modeling to be advisable with kriging. Joseph (2006) and Joseph, Hung, and Sudjianto (2008) recently proposed methods for avoiding this reversion to the trend and selecting a trend model that works well with kriging.

In the center panel, using a quadratic trend still makes the MLE $\widehat{\boldsymbol{\theta}}$ much larger (6.4 with trend vs. 0.64 without trend), corresponding to smaller correlations (74% vs. 97% between simulation outputs at adjacent design points), but the correlation is strong enough with or without the trend that kriging behaves like interpolation instead of like regression. (Indeed, kriging is no better than linear interpolation in this one-dimensional example. However, linear interpolation is not possible in higher dimension, where interpolation schemes are more complicated.) The presence of trend modeling has little impact on the quality of prediction here, a finding frequently reported by those who apply kriging to simulation output (Kleijnen 2009).

Kriging predictions improved from the left panel to the center panel, as we added design points while decreasing the effort per design point so that the total simulation budget stayed the same. The right panel of Figure 4 shows what happens if we push this too far: because it does no filtering, kriging does not cope well with significant amounts of noise from stochastic simulation. Smoothing would be more appropriate than interpolation for this data set. This is a motivation for the development of stochastic kriging, which combines smoothing and interpolation. The effort at each design point is a limiting factor on the quality of kriging metamodels of stochastic simulation. Given a fixed simulation budget, this pushes us to use experiment designs with a smaller number of design points and more effort at each. In high dimension, that can become a problem, as it becomes difficult to find an economical experiment design that enables estimation of the parameters, and the few design points become very far apart, causing poor interpolation. Stochastic kriging is intended to overcome this problem.

## 5 STOCHASTIC KRIGING

We know that stochastic simulation outputs have random errors with heterogeneous variances and, if common random numbers are used, correlations among themselves. We call the resulting uncertainty about the response surface *intrinsic* uncertainty, because it is inherent to stochastic simulation. By contrast, the *extrinsic* uncertainty of the random field in kriging is not a property of the simulation model itself, but a description of our uncertainty about the response surface at a point where we have not run a simulation. The purpose of stochastic kriging is to handle intrinsic and extrinsic uncertainty in metamodeling of stochastic simulations, thus enabling better prediction when both sources of uncertainty are non-negligible.

There is more than one way to try to achieve this goal. For example, Siem and den Hertog (2007) study the sensitivity of kriging to errors in simulation output and develop kriging methods that are less sensitive to these errors. Their approach has something in common with methods such as robust regression (Weisberg 1985, § 11.1) that seek to provide good but suboptimal performance despite violation of an assumption—in this case, kriging's assumption that the simulation output exactly equals the response surface. Stochastic kriging takes a different approach: to enrich the kriging model by providing a statistical framework that handles the errors in the simulation output.

Actually, one such method, kriging with measurement error or a "nugget effect," is standard in spatial statistics (Cressie 1993, Ch. 3). It is based on the assumption

$$Y(x) = y(x) + \varepsilon(x), \quad y(x) = \mathrm{Y}(x), \quad \mathrm{Y} \sim \mathrm{GRF}(\mu, \sigma^2), \quad \mu(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \sigma^2(x,x') = \tau^2 r(x-x';\boldsymbol{\theta}), \quad \varepsilon \sim \mathrm{WN}(v), \tag{10}$$

where $\mathrm{Y}$ and $\varepsilon$ are independent. That is, the output is the sum of the extrinsic Gaussian random field $\mathrm{Y}$, whose realization is the response surface $y$ that we aim to predict, and an independent intrinsic white-noise random field $\varepsilon$, representing measurement error in statistics, or Monte Carlo error in stochastic simulation. In spatial statistics, it is typical to regard the intrinsic variance $v$ as unknown and to estimate it based on the data $\boldsymbol{Y}$, just as in OLS regression: cf. Equation (1). That is, kriging with measurement error is an extension of kriging in which the parameters to be chosen are $\boldsymbol{\beta}$, $\tau^2$, $\boldsymbol{\theta}$, and also $v$. For $v = 0$, Equation (10) becomes kriging, and for $\tau^2 = 0$, it becomes OLS regression. Unlike in regression, the magnitude of the intrinsic variance $v$ is important and affects predictions. This is because the relative magnitudes of $v$ and $\tau^2$ matter: the greater $v/\tau^2$, the more kriging with measurement error will regard deviations from trend as noise to be filtered out, and the smoother the predictions will be. This suggests that it would be better to use any available information about intrinsic variance, which we typically have in stochastic simulation, than to choose it by maximum likelihood estimation. Because Equation (10) is a misspecified model (see Section 7.1), likelihood maximization can not be relied upon to choose $v$ well. Furthermore, maximizing likelihood over more parameters can result in worse predictions using the MLEs, as shown by the discussion of the left plot of Figure 4, where better predictions result by forcing $\beta_1 = \beta_2 = 0$. According to Kleijnen (2008, 2009), kriging with measurement error has not been widely successfully applied to metamodeling of stochastic simulation.

Stochastic kriging implements the suggestion that we should base the intrinsic covariance structure on the estimated covariances of stochastic simulation outputs. We focus on the case of independent sampling, so the variances can be estimated by sample variances if there are enough independent replications, and by standard methods (Glynn 2006, Goldsman and Nelson 2006) for single-run steady-state simulation. Ankenman, Nelson, and Staum (2008) show that typically there is only a slight loss of accuracy in prediction due to estimating intrinsic variances instead of knowing them. They also describe how the intrinsic covariance function can itself be metamodeled as a random field, much as stochastic kriging is about metamodeling the mean function. This may be particularly helpful if common random numbers are used at many design points, in which case there are a large number of covariances to estimate.

Stochastic kriging is based on the assumption

$$Y(x) = y(x) + \varepsilon(x), \quad y(x) = \mathrm{Y}(x), \quad \mathrm{Y} \sim \mathrm{GRF}(\mu, \sigma^2), \quad \mu(x) = \boldsymbol{b}(x)\boldsymbol{\beta}, \quad \sigma^2(x,x') = \tau^2 r(x - x'; \boldsymbol{\theta}), \quad \varepsilon \sim \mathrm{GRF}(0,c), \tag{11}$$

where $\mathrm{Y}$ and $\varepsilon$ are independent. Focusing on independent sampling, we further assume $c(x,x') = 0$ if $x \neq x'$. We plug in the sample variance $\mathscr{S}^2(x_i)$ of $Y_1(x_i), \ldots, Y_{n_i}(x_i)$ in place of $v(x) = c(x,x)$. *Thus stochastic kriging is analogous to EWLS, while kriging with measurement error is analogous to OLS.* The stochastic kriging prediction is similar to Equation (8) for kriging. Given the mean and covariance function of the GRF $\mathrm{Y}$, the prediction is

$$\widehat{y}(x) = \widehat{\mathrm{Y}}(x) = \boldsymbol{b}(x)\boldsymbol{\beta} + \boldsymbol{\sigma^2}(x)(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}) = \boldsymbol{b}(x)\boldsymbol{\beta} + \boldsymbol{r}(x)(\boldsymbol{R} + \boldsymbol{C}/\tau^2)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}) \tag{12}$$

where $\boldsymbol{\sigma^2}(x)$ is a row vector whose $i$th element is $\sigma^2(x,x_i)$, $\boldsymbol{\Sigma}_{hi} = \sigma^2(x_h,x_i)$, the intrinsic covariance matrix $\boldsymbol{C}$ of the output at the design points is given by $\boldsymbol{C}_{hi} = c(x_h,x_i)$, $\boldsymbol{Y}$ is the vector of output at the design points, and $\boldsymbol{B}$ has $i$th row $\boldsymbol{B}_{i\cdot} = \boldsymbol{b}(x_i)$. Given a stationary extrinsic covariance function $\sigma^2 = \tau^2 r$, we divide by the extrinsic variance $\tau^2$ (so $\boldsymbol{r}(x) = \boldsymbol{\sigma^2}(x)/\tau^2$ and $\boldsymbol{R} = \boldsymbol{\Sigma}/\tau^2$) and find that the matrix $\boldsymbol{C}/\tau^2$ of noise-to-signal ratios plays an important role. The vector $\boldsymbol{r}(x)(\boldsymbol{R} + \boldsymbol{C}/\tau^2)^{-1}$ can be interpreted as weights for interpolation; as in kriging, they may sum to less than one.

Letting $\boldsymbol{C} \to 0$ or $\tau^2 \to \infty$ causes Equation (12) to approach Equation (8): *stochastic kriging becomes like kriging when intrinsic variance is negligible compared to extrinsic variance.* Letting $\tau^2 \to 0$ or the intrinsic variances all go to infinity causes Equation (12) to approach $\boldsymbol{b}(x)\boldsymbol{\beta}$. Suppose for a moment that the extrinsic covariance function is known, but $\boldsymbol{\beta}$ is unknown and estimated by its MLE

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{B}^\top(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^\top(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}\boldsymbol{Y}. \tag{13}$$

As $\tau^2 \to 0$ or as the intrinsic variances all go to infinity, the MLE $\widehat{\boldsymbol{\beta}}$ approaches Equation (2), and the prediction $\widehat{\mathrm{Y}}(x)$ approaches $\boldsymbol{b}(x)\widehat{\boldsymbol{\beta}}$. Thus, when simulation uses independent sampling, intrinsic variances are estimated, and coefficients in the trend are chosen by maximum likelihood, *if extrinsic variance is negligible compared to intrinsic variance, stochastic kriging becomes like EWLS regression.* (It would become like GLS regression if common random numbers are used.)

The top row of Figure 5 uses the simulation output that appeared in Figure 4 to compare stochastic kriging with no trend to kriging with no trend and EWLS regression with a quadratic trend. The exponential and Gaussian correlation functions are given by $r_0(t;\theta) = \exp(-\theta t^p)$ with $p = 1$ or 2, respectively. With $p = 1$, stochastic kriging makes predictions that are much like those of kriging in the top left panel. The predictions have humps which are due in part to interpolating when the design points are far apart, in part to reversion to the mean. They are not present in the metamodels built with denser designs because the design points are closer together and the estimated correlations are larger. Moving from left to right, we see that the higher the output variance is at a point, the more smoothing stochastic kriging performs there. The smoothing effect occurs when intrinsic variance is locally substantial compared to extrinsic variance. This is why the smoothing effect is greatest for the dense design (top right panel), which has the smallest sample sizes per design point and largest intrinsic variances. We do not do kriging with $p = 2$ because the correlation matrix becomes nearly singular (Section 7.2). With $p = 2$, stochastic kriging is able to perform much more local smoothing, even where intrinsic variance is small; in the top right panel, its predictions are somewhat similar to those of EWLS regression. *One advantage of stochastic kriging is that it can yield good metamodels for a broad range of experiment designs*, including some where kriging does badly (top right panel) and some where regression does badly (top left panel). However, a disadvantage of stochastic kriging is that it is subject to the same pathology as EWLS regression: it virtually ignores simulation outputs whose estimated intrinsic variances are large compared to the estimated extrinsic variance $\widehat{\tau}^2$. See Equation (12) and subsequent discussion of $\tau^2$. This effect indicates a misspecification of (11) that will be addressed in Section 7.1.

The bottom row of Figure 5 deals with Example 2. The experiment designs are $k = 6$ design points and $n = 90$ replications (left), $k = 9$ and $n = 60$ (center), and $k = 12$ and $n = 45$ (right). We do not perform EWLS regression because the intrinsic variance is zero at some design points, leading to infinite weights. The intrinsic variance is small compared to the variability among the simulation outputs, so again we see that for $p = 1$, kriging and stochastic kriging make similar predictions. Because the intrinsic variance is relatively small and largest is in the center of $\mathscr{X}$ (not at the left edge as in Example 1), stochastic kriging makes good predictions throughout the region. Again the predictions are better for $p = 2$, which is appropriate for smooth response surfaces such as we have here (Santner, Williams, and Notz 2003).

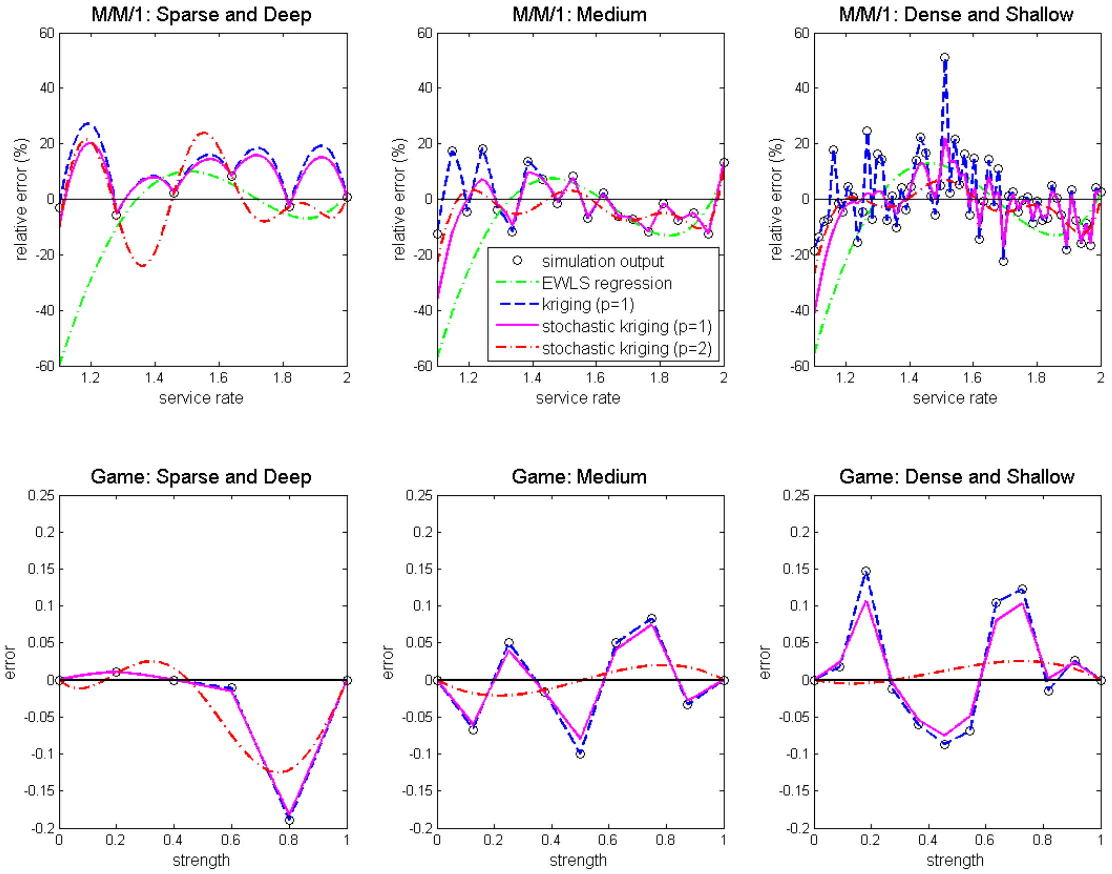

Figure 5: Metamodels of the M/M/1 simulation (Example 1, top row) and of a simulation of a simple game (Example 2, bottom row). The columns depict the results of simulation experiments with different experiment designs. The regression trend in expected waiting time is quadratic in service rate, but the kriging methods do not use a trend. The exponential and Gaussian correlation functions have $p = 1$ and $p = 2$ respectively.

**Example 2** *Players A and B toss coins independently. Player A's coin comes up heads with probability x, player B's with probability* $1-x$. *A player wins $1 by tossing heads when the opponent tosses tails. The response surface is player A's expected winnings:* $y(x) = 2x - 1$. *The intrinsic variance is* $v(x) = x^2 + (1-x)^2$.

## 6 HOW TO DO STOCHASTIC KRIGING

There are two steps in using stochastic kriging to build a metamodel from simulation output: first, use the output to choose parameters for the GRF; second, use this GRF to make predictions. We will make software to perform these tasks available at `www.stochastickriging.net`. We describe prediction first because it is simpler, and our method of choosing parameters by maximum likelihood estimation relies on the prediction algorithm.

### 6.1 Prediction

We have implemented Equation (12) via the Cholesky factorization $\boldsymbol{\Sigma} + \boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower-triangular matrix. Numerical problems arise in Cholesky factorization (or inversion) of a nearly singular matrix. Sections 6.2 and 7.2 contain advice about how to prevent $\boldsymbol{\Sigma} + \boldsymbol{C}$ from being nearly singular. Define the residuals $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}$. Then $\boldsymbol{Z} = \boldsymbol{L}^{-1}\tilde{\boldsymbol{Y}}$ is a standard $k$-variate normal vector. That is, $\boldsymbol{L}^{-1}$ is the matrix that transforms the zero-mean residuals $\tilde{\boldsymbol{Y}}$, which have covariance matrix $\boldsymbol{\Sigma}+\boldsymbol{C}$, into a vector $\boldsymbol{Z}$ of independent standard normal random variables, which can be interpreted as the underlying randomness (extrinsic and intrinsic) that caused the residuals. A procedure for prediction as in Equation (12) is:

1. From the simulation, get output vector $\boldsymbol{Y}$ and estimated intrinsic covariance matrix $\boldsymbol{C}$; compute residuals $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta}$.
2. Compute the extrinsic covariance matrix $\boldsymbol{\Sigma}$.
3. Compute the lower-triangular Cholesky factor $\boldsymbol{L}$ of $\boldsymbol{\Sigma}+\boldsymbol{C}$.
4. Solve the system of linear equations $\boldsymbol{L}\boldsymbol{Z} = \tilde{\boldsymbol{Y}}$ to get $\boldsymbol{Z} = \boldsymbol{L}^{-1}\tilde{\boldsymbol{Y}}$.
5. Solve the system of linear equations $\boldsymbol{L}^\top\boldsymbol{Q} = \boldsymbol{Z}$ to get $\boldsymbol{Q} = (\boldsymbol{\Sigma}+\boldsymbol{C})^{-1}\tilde{\boldsymbol{Y}}$.
6. Predict $\widehat{y}(x) = \boldsymbol{b}(x)\boldsymbol{\beta} + \boldsymbol{\sigma}^2(x)\boldsymbol{Q}$.

There are three parts: output analysis in Step 1, computation of $\boldsymbol{Q}$ in Steps 2–5, and prediction at $x$ in Step 6. With independent sampling, which makes $\boldsymbol{C}$ diagonal, Step 1 is fast. Cholesky factorization in the second part dominates the computational complexity of the procedure: it is $\mathscr{O}(k^3)$ in the number $k$ of design points. Section 7.2 discusses how to reduce the computational cost. The third part, the only part that has to be done after learning the prediction point $x$, is very fast.

### 6.2 Choosing Parameters of the Gaussian Random Field

We discuss likelihood maximization for choosing the parameters. For other methods of choosing the parameters in kriging, see Cressie (1993) or Santner, Williams, and Notz (2003). Given a choice of $\tau^2$ and $\boldsymbol{\theta}$, the value $\widehat{\boldsymbol{\beta}}(\tau^2,\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ that maximizes likelihood is given by Equation (13). Therefore the MLEs $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\theta}}$ maximize the *profile log-likelihood* given by

$$\mathscr{L}(\tau^2,\boldsymbol{\theta}) = -\frac{1}{2}\left(k\ln(2\pi)+\ln(|\boldsymbol{\Sigma}+\boldsymbol{C}|)+\tilde{\boldsymbol{Y}}^\top(\boldsymbol{\Sigma}+\boldsymbol{C})^{-1}\tilde{\boldsymbol{Y}}\right) = -\frac{1}{2}k\ln(2\pi)-\ln(|\boldsymbol{L}|)-\frac{1}{2}\|\boldsymbol{Z}\|^2, \tag{14}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\tau^2,\boldsymbol{\theta})$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\beta}}(\tau^2,\boldsymbol{\theta})$. Every evaluation of $\mathscr{L}$ involves performing Steps 2–5 of the procedure in Section 6.1. We minimize $-\mathscr{L}$ using MATLAB's nonlinear solver `fmincon`. It seems desirable to provide the solver with the gradient and Hessian of $\mathscr{L}$.

We have found it helpful to start by normalizing the design points, whose different components may have very different variabilities: for example, one component may represent throughput ranging from 10,000 to 20,000 units per day, and another component may represent a buffer size ranging from 10 to 100. Normalization makes it easier to choose tolerances for the nonlinear solver and to handle the gradient with respect to $\boldsymbol{\theta}$ numerically. We normalize the design points by transforming each component separately, mapping $x_{ih}$ to $\left(x_{ih} - \min_{j=1,\dots,k} x_{jh}\right) / \left(\max_{j=1,\dots,k} x_{jh} - \min_{j=1,\dots,k} x_{jh}\right)$ for each design point $x_i$ and component $h$, so that the design points fall in the unit hypercube $[0,1]^d$. This affine transformation merely changes the scale of each component of $\boldsymbol{\theta}$; it does not change the metamodel built with the likelihood-maximizing parameters.

The nonlinear solver requires a starting point. We present a heuristic method for choosing one. We aim not to start at a point far from the optimum where any components of the gradient are very small. For example, if $\tau^2$ is too large, $\partial\mathscr{L}/\partial\tau^2$ is small. We choose the initial value of $\tau^2$ to be the variance of the residuals of OLS regression, $\mathrm{Var}[Y - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top Y]$. To choose the initial value of $\boldsymbol{\theta}$, we focus on spatial correlation functions of product form, as in Equation (9). In most such models, if $\theta_h$ is too small or too large, $\partial\mathscr{L}/\partial\theta_h$ is small because the matrix $\partial\boldsymbol{R}/\partial\theta_h$ of correlations' sensitivities to $\theta_h$ is small. To get a moderate value for each $\theta_h$, we let $\bar{x}_h$ be an average distance in dimension $h$ among design points, such as $\bar{x}_h = \sum_{i=1}^k \sum_{j=i+1}^k |x_{ih} - x_{jh}|/(k(k-1)/2)$, and we start with the value of $\theta_h$ that solves the equation $0.5 = (r_0(\bar{x}_h;\theta_h))^d$.

The optimization should also include some constraints. Because $\partial\mathscr{L}/\partial\tau^2$ may be negative at $\tau^2 = 0$, it is important to include the constraint $\tau^2 \geq 0$. Many spatial correlation models require $\boldsymbol{\theta} \geq 0$. Numerical instability in Cholesky factorization results if $\boldsymbol{\Sigma}+\boldsymbol{C}$ is nearly singular, which happens if some rows of $\boldsymbol{\Sigma}$ are nearly collinear due to extremely high spatial correlation between nearby points and those rows of $\boldsymbol{C}$ are nearly zero or nearly collinear. This can be avoided by experiment design, as discussed in Section 7.2. It can also be avoided by preventing spatial correlations among distinct design points from growing too large during likelihood maximization. In many spatial correlation models, this can be achieved by imposing an artificial constraint such as $\boldsymbol{\theta} \geq \vartheta_0$ or $\sum_{h=1}^d \theta_h \geq \vartheta_0$ for some small $\vartheta_0$.

## 7 CONCLUSIONS AND FUTURE DIRECTIONS

Limited initial experimentation has suggested the following tentative conclusions:

- Unlike regression, kriging and stochastic kriging can make good predictions without a well-specified trend model.

- Because stochastic kriging can tolerate much more Monte Carlo noise than kriging, it permits the use of smaller effort per design point and more design points.
- Like kriging, stochastic kriging can be slower than regression when there are many design points.
- Like kriging but unlike regression, stochastic kriging is vulnerable to misspecification of the GRF.

### 7.1 Misspecification and Validation

One should validate a metamodel before using it, to avoid making severely inaccurate predictions. Bastos and O'Hagan (2008) propose methods for validating metamodels based on GRFs, and Kleijnen (2008) discusses validation in metamodeling of stochastic simulation. An invalid metamodel can result because the experiment design did not provide enough data, because of over-fitting, or because of misspecification. Over-fitting is a danger when there are too few design points relative to the number of parameters being chosen, but it can be mitigated by methods such as factor screening and penalized optimization (Fang, Li, and Sudjianto 2006, Kleijnen 2008). Welch et al. (1992) propose something analogous to factor screening for correlation coefficients $\boldsymbol{\theta}$ in kriging: start by assuming that all correlation coefficients are the same, and one-by-one allow some dimension to have a different correlation coefficient if it greatly increases the likelihood. We conjecture that stochastic kriging also mitigates the danger of over-fitting, compared to kriging, by allowing for a larger number of design points given a fixed computational budget.

Stein (1999) discusses misspecification in kriging extensively, primarily considering what happens when the data is generated by a GRF with a stationary covariance function but kriging is performed under incorrect assumptions about the GRF's mean and covariance function. From a frequentist perspective, the response surface is not random, so the conceptual framework for misspecification of kriging in simulation metamodeling is less straightforward than in spatial statistics. O'Hagan (2006) provides a tutorial on the Bayesian perspective. Loosely speaking, we can say that there may be a problem with misspecification if the response surface $y$ has very low likelihood given the GRF assumption. For one, the structure of the spatial correlation function may be misspecified (see Section 4).

In simulation metamodeling practice, it seems that assuming that the covariance function is stationary can be a damaging form of misspecification. Fortunately, the assumption of stationary covariance can be lifted, just as it was possible to lift the assumption that the mean function is stationary (i.e., constant) to model the trend $\boldsymbol{b}(\cdot)\boldsymbol{\beta}$. Some major approaches are exemplified by the following recent papers (see also references therein). Paciorek and Schervish (2006) work with a non-stationary covariance function and propose methods for local estimation of covariance parameters. Xiong et al. (2007) show how to create a spatial transformation such that it is more acceptable to assume stationarity of the covariance function in the transformed space. Gramacy and Lee (2008) develop a statistical technique for partitioning the space into regions in which kriging metamodeling is performed separately. This approach is attractive insofar as local metamodeling is easier than global metamodeling, but the global metamodel formed by pasting the local metamodels together is generally discontinuous, which may be desirable or undesirable depending on the application and the use to which the metamodel will be put.

Distributional assumptions are another potential source of misspecification in stochastic kriging. Kriging treats the deviations of the response surface from the trend as Gaussian for the sake of tractability. Stochastic kriging treats the intrinsic simulation error as Gaussian, which is typically a good assumption. However, in some situations where the simulation effort $n$ at a design point is small, alternative assumptions could be better: for example, the simulation output is binomially distributed if the output of each replication is an indicator function, e.g. whether 99% of all calls were answered within 5 minutes. One may avoid these Gaussian assumptions by using the results of Diggle, Tawn, and Moyeed (1998), who generalized kriging in the same way that the standard linear model was generalized to the GLM (Section 3). For example, if $\ln \mathsf{Y}$ is modeled as a GRF, $\mathsf{Y}(x)$ has a lognormal distribution. Although we did not explore it in Section 3, another feature of GLMs is that they permit the error to be assigned some distributions other than Gaussian.

### 7.2 Experiment Design

Different experiment designs, even with the same total simulation effort, can lead to quite different kriging and stochastic kriging metamodels (Figure 5). There is a substantial body of knowledge about experiment design for kriging in metamodeling of deterministic and stochastic simulation (Fang, Li, and Sudjianto 2006, Kleijnen 2008, Santner, Williams, and Notz 2003), but very little is known yet about experiment design specifically for stochastic kriging. However, we can indicate some differences between experiment design for metamodeling of stochastic simulation by stochastic kriging as opposed to kriging.

With kriging, we simply know that simulation output variance should not be too large, so we use fairly large simulation effort at all design points, or use adaptive effort to reach a fixed target for output variance (Kleijnen 2008, § 5.4). Stochastic kriging takes into account how simulation effort changes the variance of the simulation output and thus impacts the predictions and their accuracy. This introduces a new aspect of experiment design for metamodeling of stochastic simulation: for example, it may be preferable to allocate greater simulation effort to design points that are near the center of the domain $\mathscr{X}$ (Ankenman, Nelson, and Staum 2008, § 3.3). As in kriging, there is great value to sequential methods of experiment design: after running some simulations, decisions about the allocation of further simulation effort are made. There is more flexibility for sequential methods to exploit with stochastic kriging than with kriging, because it is possible to have low simulation effort at design points initially and then to perform more simulation at design points selectively. Ankenman, Nelson, and Staum (2008) discuss doing this for global metamodeling. Liu and Staum (2009) have a procedure that does this for a financial application that demands accuracy of the metamodel only in a subset of $\mathscr{X}$ that is initially unknown and must be discovered.

Those papers rely on a key tool for sequential experiment design in kriging, the predictive variance of $\mathsf{Y}(x)$, which expresses how much uncertainty there is in predicting $\mathsf{Y}(x)$ given the simulation output. The value of a planned experiment can be assessed via the expected decrease in predictive variance due to observing the data that will be simulated. Based on Equation (11), Ankenman, Nelson, and Staum (2008) give a formula for the predictive variance of stochastic kriging. Unfortunately, this formula is even more sensitive to misspecification of the GRF than predictions (Stein 1999). Consequently, Kleijnen (2008) instead relies on bootstrapping to assess predictive variance.

Another way in which experiment design differs for kriging and stochastic kriging is that the ability of stochastic kriging to tolerate larger output variance permits smaller simulation effort per design point. (In steady-state simulation, bias may limit how low the effort can be.) Thus, given a fixed computational budget, stochastic kriging permits a larger number $k$ of design points than kriging. This can be advantageous because it enables better estimation (cf. the predictions of kriging with quadratic trend in the left vs. the center plots of Figure 4) and improves the accuracy of interpolation.

However, due to the $\mathscr{O}(k^3)$ computational cost discussed in Section 6.1, $k$ should not be taken too large with this algorithm. There has recently been a lot of work in spatial statistics on faster procedures for kriging with many design points, some of which can be adapted to the setting of stochastic kriging: see Cressie and Johannesson (2008) on fixed rank kriging, Kaufman, Schervish, and Nychka (2008) on covariance tapering, and references therein. The tree technique of Gramacy and Lee (2008) may help because it has the side effect of reducing the number of design points per GRF.

Another computational issue with implications for experiment design is that numerical problems related to near-singularity of the covariance matrix arise in Cholesky decomposition if there are design points whose simulation outputs are too highly correlated. This happens in ordinary kriging when design points are too close to each other, making extrinsic correlation large. It can also happen in stochastic kriging if intrinsic correlation is large too, or if intrinsic variance is negligible. The lesson for experiment design with stochastic kriging is that you can put design points wherever you want, but do not give points that are close to each other sample sizes so large that intrinsic variance is negligible at both.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. DMI-0555485 and DMI-0900354. The author thanks Xi Chen, Ming Liu, Barry Nelson, and Wei Xie for comments, corrections, and references.

## REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2008. Stochastic kriging for simulation metamodeling. *Operations Research*. Forthcoming.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation: Algorithms and analysis*. New York: Springer-Verlag.

Barton, R. R., and M. Meckesheimer. 2006. Metamodel-based simulation optimization. In *Simulation*, ed. S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, Chapter 18, 535–574. Amsterdam: Elsevier.

Bastos, L. S., and A. O'Hagan. 2008, January. Diagnostics for Gaussian process emulators.

Borogovac, T., and P. Vakili. 2008. Control variate technique: a constructive approach. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 320–327. Piscataway, N. J.: IEEE Press.

Cheng, R. C. H., and J. P. C. Kleijnen. 1999. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Operations Research* 47 (5): 762–777.

Cressie, N., and G. Johannesson. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B* 70:209–226.

Cressie, N. A. C. 1993. *Statistics for spatial data*. New York: John Wiley & Sons.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. Model-based geostatistics. *Journal of the Royal Statistical Society, Series C* 47 (3): 299–350.

Fang, K.-T., R. Li, and A. Sudjianto. 2006. *Design and modeling for computer experiments*. Boca Raton: Chapman & Hall/CRC.

Fu, M. C. 2008. What you should know about simulation and derivatives. *Naval Research Logistics* 55 (8): 723–736.

Glynn, P. W. 2006. Simulation algorithms for regenerative processes. In *Simulation*, ed. S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, Chapter 16, 476–500. Amsterdam: Elsevier.

Goldsman, D., and B. L. Nelson. 2006. Correlation-based methods for output analysis. In *Simulation*, ed. S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, Chapter 15, 455–475. Amsterdam: Elsevier.

Gramacy, R. B., and H. K. H. Lee. 2008. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* 103 (483): 1119–1130.

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

Joseph, V. R., Y. Hung, and A. Sudjianto. 2008. Blind kriging. *Journal of Mechanical Design* 130 (3).

Kaufman, C. G., M. J. Schervish, and D. W. Nychka. 2008. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* 103 (484): 1545–1555.

Kim, S., and S. G. Henderson. 2008. The mathematics of continuous-variable simulation optimization. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 122–132. Piscataway, N. J.: IEEE Press.

Kleijnen, J. P. C. 2008. *Design and analysis of simulation experiments*. New York: Springer-Verlag.

Kleijnen, J. P. C. 2009. Kriging metamodeling in simulation: a review. *European Journal of Operational Research* 192:707–716.

Liu, M., and J. Staum. 2009, March. Estimating expected shortfall with stochastic kriging.

McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*. 2nd ed. New York: Chapman & Hall.

Morris, M. D., T. J. Mitchell, and D. Ylvisaker. 1993. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics* 35 (3): 243–255.

O'Hagan, A. 2006. Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety* 91 (10-11): 1290–1300.

Paciorek, C. J., and M. J. Schervish. 2006. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 17:483–506.

Qian, P. Z. G., H. Wu, and C. F. J. Wu. 2008. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* 50 (3): 383–396.

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

Siem, A. Y. D., and D. den Hertog. 2007, August. Kriging models that are robust with respect to simulation errors. Discussion paper 2007-68, CentER, Tilburg University.

Stein, M. L. 1999. *Interpolation of spatial data: Some theory for kriging*. New York: Springer-Verlag.

Weisberg, S. 1985. *Applied linear regression*. 2nd ed. New York: John Wiley & Sons.

Welch, W. J., R. T. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris. 1992. Screening, predicting, and computer experiments. *Technometrics* 34 (1): 15–25.

Xiong, Y., W. Chen, D. Apley, and X. Ding. 2007. A non-stationary covariance-based Kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering* 71:733–756.

## AUTHOR BIOGRAPHY

**JEREMY STAUM** is Associate Professor of Industrial Engineering and Management Sciences and holds the Pentair-Nugent Chair at Northwestern University. He is a fellow of the FDIC Center for Financial Research. He coordinated the Risk Analysis track of the 2007 Winter Simulation Conference and serves as department editor for financial engineering at *IIE Transactions* and as an associate editor at *Operations Research*. His website is <users.iems.northwestern.edu/ staum>