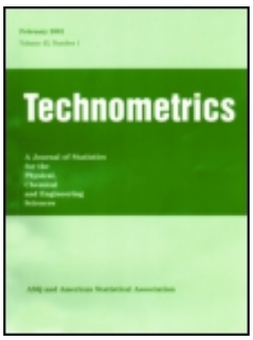

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Building Accurate Emulators for Stochastic Simulations via Quantile Kriging

Matthew Plumlee & Rui Tuo

To cite this article: Matthew Plumlee & Rui Tuo (2014) Building Accurate Emulators for Stochastic Simulations via Quantile Kriging, Technometrics, 56:4, 466-473, DOI: 10.1080/00401706.2013.860919

To link to this article: https://doi.org/10.1080/00401706.2013.860919

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Building Accurate Emulators for Stochastic Simulations via Quantile Kriging

Matthew PLUMLEE

Georgia Institute of Technology
North Avenue NW
Atlanta, GA 30332
(*mplumlee@gatech.edu*)

Rui TUO

Chinese Academy of Sciences
100864 Beijing, China
(*tuorui@amss.ac.cn*)

Computer simulation has increasingly become popular for analysis of systems that cannot be feasibly changed because of costs or scale. This work proposes a method to construct an *emulator* for stochastic simulations by performing a designed experiment on the simulator and developing an emulative distribution. Existing emulators have focused on estimation of the *mean* of the simulation output, but this work presents an emulator for the *distribution* of the output. This construction provides both an explicit distribution and a fast sampling scheme. Beyond the emulator description, this work demonstrates the emulator's efficiency, that is, its convergence rate is the asymptotically optimal among all possible emulators using the same sample size (under certain conditions). An example of its practical use is demonstrated using a stochastic simulation of fracture mechanics. Supplementary materials for this article are available online.

KEY WORDS: Computer experiments; Gaussian process; Metric entropy; Quantile regression; Reproducing kernel Hilbert spaces; Simulation experiments.

1. INTRODUCTION

Computer simulation is widely used to measure the performance of systems in the presence of stochastic behavior. Typically, the simulation has a collection of inputs which represent a variety of unknown or controllable aspects of the system. However, these simulations can be computationally expensive to run in fine-mesh or large-scale simulation environments. The investment in the development of computer models can be lost if, for example, a large number of alternative inputs need to be investigated or the desired analysis requires repeated evaluations over long periods of time where computer clusters may be unavailable. For example, take the propagation of cracks in metals, where the stochastic nature of grain formation creates uncertainty in fracture growth rates (Stephens and Fuchs 2001). The proliferation of increasingly complex numerical algorithms for fatigue analysis necessitates a limited sample size (see Section 5.1 for examples). However, implementation scenarios involving online condition monitoring, for example, Ray and Tangirala (1996), require a large number of evaluations with a limitless sample size.

This work proposes a method to *emulate* the stochastic simulation with a simple stochastic model. The emulator of a stochastic simulation provides two important constructions: (1) an explicit functional form of the distribution and (2) a fast sampling scheme. The emulator can then be integrated into analysis software (e.g., spreadsheet environments), which allows for timely results from investigations such as what-if scenarios and uncertainty quantification. An emulator is created by establishing an emulative distribution of the simulation output based on observations from an experiment. The An emulative distribution is based on a stochastic model representing the simulation output, termed the *metamodel*. As has been shown in multiple disciplines, including geostatistics (Matheron 1963; Diggle and Ribeiro 2007) and analysis of deterministic computer code (Sacks et al. 1989; Santner, Williams, and Notz 2003), random field metamodels often offer superior representation of underlying continuous functions compared to low order polynomial metamodels (Barton 1998). The use of these random field metamodels is commonly referred to as *kriging*.

While previous attempts using random field metamodels have focused on the *mean* of the simulation output, an emulator for the simulation's *stochastic behavior* is often needed. Ankenman, Nelson, and Staum (2010) described the case when the variance of the output significantly changes with alterations to the inputs, an important concern in stochastic simulations. However, the use of the traditional random field metamodel as in Kleijnen (2007); Ankenman, Nelson, and Staum (2010); and Picheny et al. (2013) is inadequate to provide emulative distributions due to a normality assumption on the stochastic behavior of the simulation output. The popular technique of model-based geostatistics (Diggle and Ribeiro 2007) and similar methods (Rigby and Stasinopoulos 2005; Henderson et al. 2009) addresses normality concerns when the distribution of the output is in a parametric class (e.g., exponential). However, parametric assumptions often do not have the power to address the complex distributions that can result from simulations, for example, Sze (1984) discussed bimodal cases. In our experience, the previously developed methods prove powerful when the respective assumptions hold, but there exist cases when a single parametric class for all inputs is not a reasonable assumption. For example,

TECHNOMETRICS, November 2014, VOL. 56, NO. 4
DOI: 10.1080/00401706.2013.860919
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/tech*.

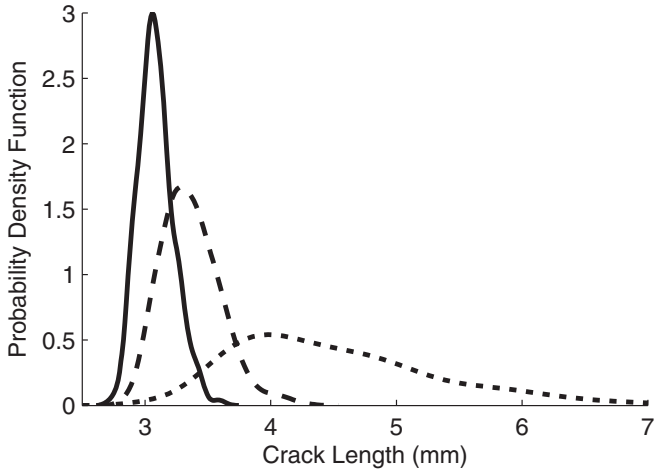

Figure 1. Empirical distributions of crack lengths after 2000 cycles with a stress ratio of 0 (solid line), 0.25 (long dashes), and 0.5 (short dashes).

take the stochastic modification of the Forman equation, which is a general model for the growth rate of fractures based on a stress ratio (details are discussed in Section 5.1). The goal considered here is the prediction of the distribution of the crack size after 2000 cycles with an initial size of 2.54 mm. Figure 1 shows the distribution of the simulation output. When the stress ratio is nearly zero, an approximately Gaussian behavior results. As the stress ratio increases, this structure breaks down, indicating that previously developed techniques cannot be used.

This article describes the development of emulators through a framework termed *quantile kriging*, which allows for nonparametric representations of the stochastic behavior of the output. The framework consists of conducting a designed experiment with replications at different sets of inputs and then establishing an emulative distribution for all inputs. The emulator comes with both an explicit form and an assoicated a fast sampling scheme.

This paper also studies the asymptotic properties of this emulator which communicate valuable insights. For example, experiments consisting of replications at sets of different inputs are nearly universally accepted among users of simulations (Ankenman, Nelson, and Staum 2010), but the rational is not always justified. We demonstrate, under certain regulatory conditions, a result that can be summarized as follows (see Section 4):

> *By using an experiment that has the appropriate ratio of replications to sets of different inputs, we can achieve an optimal rate of convergence.*

To the authors' knowledge, this is the first result of this type for stochastic emulators.

The basic idea of the proposed framework is to estimate the underlying *quantiles* of the distribution. After discussing the modeling strategy in Section 2, we propose a method to develop emulative distributions in Section 3. Sections 4 and 5 demonstrate the advantages of this framework by investigating the asymptotic efficiency and two illustrations, respectively. Section 6 briefly discusses some conclusions, comparisons to other work, and possible extensions of this work.

2. SIMULATION METAMODELING

As mentioned in Section 1, emulators are traditionally developed using random field metamodels (Sacks et al. 1989; Santner, Williams, and Notz 2003) which provide the ability to model simulation output without the restrictive linear or low-order polynomial assumptions. Let x represent a single input in the d dimensional input space. The basic idea of the traditional metamodel for stochastic simulations (Kleijnen 2007; Ankenman, Nelson, and Staum 2010; Picheny et al. 2013) assumes that the output is the sum of a deterministic, but unknown, mean $M(x)$ and a random variable $\varepsilon(x)$ representing the stochastic behavior of the simulation, that is,

$$Y(x) = M(x) + \varepsilon(x),$$
$$\varepsilon(x) \sim \mathcal{N}(0, \sigma^2(x)),$$

where $\sigma^2(x)$ represents the variance of the output, which is a function of the inputs. The value of $Y(x)$ represents a single draw from the simulation with inputs x. In the interest of generality, we assume only independent samples are drawn from the simulation model. Since $M(x)$ is unknown but deterministic, the framework from deterministic simulations is adopted, for example, Sacks et al. (1989), and a distributional assumption is placed on $M(x)$ that represents our uncertainty. The deterministic value of M has a prior distribution of $\mathcal{GP}(\mu(\cdot), C(\cdot,\cdot))$, where $\mathcal{GP}$ denotes a Gaussian process with a trend function $\mu(x)$ and a covariance structure $\text{cov}(M(x), M(x')) = C(x, x')$.

This approach is limited by the normality assumption on $\varepsilon(x)$, which, as mentioned in the introduction, is often invalid. Let $Q_\alpha(x)$ represent the α quantile of the distribution of $Y(x)$, that is $Q_\alpha(x) = \inf\{t : P(Y(x) \le t) \ge \alpha\}$. Here, we establish the key idea of the proposed framework: *since Q_α is an unknown function, we ought to attempt to estimate it as a function of x*. Therefore, we model the quantiles as unknown functions of x, and we further assume they are continuous. Emulators typically work by exploiting the continuity, or higher orders of differentiability, of the output (discussed in, e.g., Santner, Williams, and Notz 2003). Without any assumptions of this variety, creating emulative distributions would prove futile.

This means that if we observed two replications from the same input x, we assume their respective simulation outputs may differ, but the distribution is the same. Two replications from x and $x' \neq x$ would have *differing* distributions of the simulation output, but similar inputs (measured in distance) imply similar distributions. Therefore, in this metamodel, the aleatoric variation need not be Gaussian, but we assume that the distribution of the simulation output is continuous, meaning as $x \to x_0$, the distribution of the output at x approaches the distribution at x_0.

As an example, let the output, $Y(x)$, be the failure time for a product, which is often modeled as exponentially distributed. Define the mean of $Y(x)$ as $\nu(x) > 0$ and assume the function $\nu(x)$ is smooth. The quantiles of the exponential distribution are given by

$$Q_\alpha(x) = -\ln(1-\alpha)/\nu(x),$$

and output quantiles, are continuous as a function of x.

Under general assumptions on the distribution of the output, we can establish the continuity of the quantiles. The following

proposition demonstrates this under a broad class of assumptions (proof is located in the supplementary materials).

Proposition 1. Let F_x be defined as the cumulative probability distribution function such that $Y(x) \sim F_x$. Suppose, $F_x(y)$ is continuous with respect to x and y and $F_x(y)$ is a strictly monotonic function with respect to y, then $Q_\alpha(x)$ is a continuous function with respect to x.

However, this is not an exclusive characterization; the output $Y(x)$ does not need to be a continuous random variable. Consider the following simplified case: you flip a coin, you win x if it lands heads side up and lose x if it lands tails side up. This example is characterized by the following simple distribution:

$$F_x = 0.5\mathbb{1}\{-x \le y\} + 0.5\mathbb{1}\{x \le y\},$$

which corresponds to a discrete distribution. Here, F_x is *not* a continuous or strictly increasing function of y, and therefore does not fit the criteria listed in Proposition 1, but the quantiles, $Q_\alpha = -x + 2x\mathbb{1}\{\alpha \ge 0.5\}$, are continuous with respect to x.

3. EMULATIVE DISTRIBUTION

This section outlines the creation of emulative distributions from a designed experiment. We assume that there is an experiment that comprises n sets of inputs, denoted as $\mathcal{X} = \{x_1, \ldots, x_n\}$, with m replications, which results in a set of observations $y_1(x), y_2(x), \ldots, y_m(x)$ for each $x \in \mathcal{X}$. The choice of $\mathcal{X}$ for use with random field models has been studied in several contexts, and the authors point to Santner, Williams, and Notz (2003), and the references therein for more information. In general, the selection of space-filling Latin hypercube designs has yielded positive results.

Using this experiment, this work seeks to develop an emulative distribution for a new input $x_0 \notin \mathcal{X}$, that is, $\hat{F}_{x_0}$, which is close to the true distribution of the simulation output, $Y(x_0) \sim F_{x_0}$. After some preliminaries, the explicit emulative distribution is described (Section 3.2). Discussions of the practical matter of estimation of parameters associated with the Gaussian process model can be seen in Section 3.3. The asymptotic analysis of the framework is outlined in Section 4.

3.1 Expository Development

For simplicity, first consider the case where only a single level α exists and we observe $Q_\alpha(x)$ for each $x \in \mathcal{X}$ but Q_α is unknown for other values of the inputs. What would be a good prediction of the α quantile at other inputs, such as x_0? Since Q_α is assumed continuous, a reasonable prior is that Q_α follows a Gaussian process with mean $\mu(\cdot)$ and covariance $C(\cdot, \cdot)$. From this, a widely used estimate of the α quantile at x_0 would be

$$\mu(x_0) + \boldsymbol{\sigma}^{\mathsf{T}}(x_0)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Q}_\alpha - \boldsymbol{\mu}\right), \tag{1}$$

where $\boldsymbol{\Sigma}$ is a matrix composed of elements $C(x, x')$ for all $x, x' \in \mathcal{X}$, $\boldsymbol{\sigma}(x_0)$ is a vector composed of elements $C(x_0, x)$ for all $x \in \mathcal{X}$, $\boldsymbol{\mu}$ is a vector composed of elements $\mu(x)$ for all $x \in \mathcal{X}$ and $\boldsymbol{Q}_\alpha$ is a vector consisting of $Q_\alpha(x)$ for all $x \in \mathcal{X}$. This estimate follows directly from the extensive work in Gaussian process models, for example Santner, Williams, and Notz (2003).

The case being considered assumes $Q_\alpha(x)$ is unknown, but we can draw m replicates from the associated distribution. Thus, we replace the estimate involving $\boldsymbol{Q}_\alpha$ with one that employs vector of estimated α quantiles at each point $x \in \mathcal{X}$. The vector of estimated α quantiles is labeled $\tilde{\boldsymbol{Q}}_\alpha = [\tilde{Q}_\alpha(x_1), \ldots, \tilde{Q}_\alpha(x_n)]^{\mathsf{T}}$. Since $Q_\alpha(x)$ differs from $\tilde{Q}_\alpha(x)$ due to the stochastic nature of the simulation, we need to introduce a nugget term to the metamodel to incorporate the random difference between $Q_\alpha(x)$ and $\tilde{Q}_\alpha(x)$. This involves using a covariance function of the form $C(x, x') + \rho^2\mathbb{1}\{x = x'\}$, where ρ^2 is a scalar that represents the variation of $Q_\alpha(x) - \tilde{Q}_\alpha(x)$. Now, an estimate of $Q_\alpha(x_0)$ is given by

$$\mu(x_0) + \boldsymbol{\sigma}^{\mathsf{T}}(x_0)(\boldsymbol{\Sigma} + \rho^2\boldsymbol{I})^{-1}(\tilde{\boldsymbol{Q}}_\alpha - \boldsymbol{\mu}).$$

Because ρ is not known exactly, the choice of this value discussed in Sections 3.3 and 4.

Since the distribution of the output of a complex simulation often cannot be placed in a parametric class, we propose using the empirical quantile estimates for $\tilde{Q}_\alpha(x)$, that is,

$$\tilde{Q}_\alpha(x) = \inf\left\{t; \sum_{i=1}^{m} \mathbb{1}(y_i(x) \le t) \ge m\alpha\right\}. \tag{2}$$

Let $y_{(k)}(x)$ represent the kth order statistic from the m replications at $x \in \mathcal{X}$. Recognize that $\alpha \in [(k-1)/m, k/m)$ implies that $\tilde{\boldsymbol{Q}}_\alpha = \boldsymbol{y}_{(k)}$ where $\boldsymbol{y}_{(k)}$ is a vector of the kth order statistic from each set of inputs in the experiment, that is, $\boldsymbol{y}_{(k)} = [y_{(k)}(x_1), \ldots, y_{(k)}(x_n)]^{\mathsf{T}}$.

3.2 Explicit Emulative Distribution

We can now establish the emulative distribution by constructing a distribution with quantiles that correspond to our estimated quantiles at x_0. A simple distribution that has this property is given by

$$\hat{F}_{x_0}(y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(y \le a_i(x_0)), \tag{3}$$

where

$$a_i(x_0) = \mu(x_0) + \boldsymbol{\sigma}^{\mathsf{T}}(x_0)\left(\boldsymbol{\Sigma} + \rho^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{y}_{(i)} - \boldsymbol{\mu}\right).$$

Because the emulative distribution is a mixture point masses, sampling from it is simple and fast.

As we will show in Section 4, this estimated distribution will be close to the true distribution as the number of different inputs in the experiment and the number of replications is increased (as measured by the closeness of the quantiles). The reason for this is the number of basis functions increases as the number of replicates is increased, thus this emulative distribution can be extremely close to a variety of nonnormal distributions including heavily tailed and bimodal distributions.

3.3 Choice of C

The assumed properties of the quantiles with respect to x depend on the choice of covariance function C. While C is required to be positive definite, there is a broad array of choices for the covariance function, and the most widely used are the Matérn and Gaussian classes of stationary correlation functions. Covariance functions are typically endowed with a set of parameters,

$\boldsymbol{\theta}$, which represent properties of the response surface including lengthscale, differentiability, and the Hausdorff dimension. The parameter ρ^2 is often included in $\boldsymbol{\theta}$ because it is unknown, even though it is not explicitly a covariance parameter but an estimation parameter that should change based on the number of observations (see Section 4).

We propose estimating these parameters from the data via cross-validation criteria, which has been shown to be an effective implementation strategy for emulators, for example, Currin et al. (1991). The cross-validation criteria measure the squared prediction error if an observation or set of observations is ignored. Therefore, selection of parameters by cross-validation intuitively results in parameters with good emulative properties. Here, the predictive performance is measured by the prediction of an estimated quantile from the previous section, that is if $\alpha \in [(k-1)/m, k/m)$ then $\tilde{\boldsymbol{Q}}_\alpha = \boldsymbol{y}_{(k)}$. Let $\boldsymbol{\Sigma}$ be defined as in Section 3.2, the leave-one-out cross-validation for each observation can be quickly calculated using

$$e_{ij}(\boldsymbol{\theta}) = \frac{(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}))_j}{(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}))_{jj}}\big(\boldsymbol{y}_{(i)} - \boldsymbol{\mu}\big),$$

where $()_j$ is the jth row and $()_{jj}$ is the jth diagonal element. Therefore, we select covariance parameters as $\hat{\boldsymbol{\theta}} = \operatorname{argmin} \sum_{i=1}^{n}\sum_{i=1}^{m} e_{ij}^2(\boldsymbol{\theta})$.

4. ASYMPTOTIC EFFICIENCY

While the next section will outline an example of the practical benefits of the proposed methodology, this section will show that the proposed method is asymptotically consistent and efficient, that is, under some regulatory conditions, no other framework can do better as $n, m \to \infty$. While sample size restrictions prevent the asymptotic results from being directly used, the consistency of the prediction is critical to gauging the performance of the proposed two-stage framework.

In this section, we assume that the Gaussian process model is *stationary*, which implies that the covariance function is only a function of the distance between two sets of inputs, that is, $C(x, x') = C(x - x')$. Without loss of generality, we further assume that the observations are normalized, that is, zero mean and $C(0) = 1$. We emphasize these assumptions on C by denoting it $\Phi(h)$, $h \in \mathbb{R}^d$. Suppose the design region $x \in \Omega$ is a convex and compact subset of $\mathbb{R}^d$. Since prior distributions for surfaces such as Gaussian processes are difficult to confirm, we demonstrate that our results in a general function space. We assume that the underlying true function $Q_\alpha(x)$ lies in the reproducing kernel Hilbert space generated by Φ, denoted as $\mathcal{N}_\Phi(\Omega)$ (for more background on these function spaces, refer to Wendland 2005).

For $0 < \alpha < 1$, we assume that $Q_\alpha(x_i)$ is estimated by the empirical distribution, as seen in Equation (3), denoted as $\tilde{Q}_\alpha(x_i)$. Invoking the representer theorem (Wahba 1990; Schölkopf, Herbrich, and Smola 2001), the value of $a_{\lfloor \alpha m \rfloor}(\cdot)$, equals to the solution to the following minimization problem for some $\lambda_{m,n}^2 > 0$

$$\hat{Q}_\alpha(\cdot) = \operatorname*{argmin}_{f \in \mathcal{N}_\Phi(\Omega)} \frac{1}{n}\sum_{i=1}^{n}(\tilde{Q}_\alpha(x_i) - f(x_i))^2 + \lambda_{m,n}^2 \|f\|_{\mathcal{N}_\Phi(\Omega)}^2. \tag{4}$$

Next, we demonstrate the efficiency of $\hat{Q}_\alpha(\cdot)$ for given α under certain regularity conditions:

(A1) $x_i \overset{\text{iid}}{\sim} U(\Omega)$, the uniform distribution over Ω.

(A2) Let F_x be defined as $Y(x) \sim F_x$. For each $x \in \Omega$, there exists $\epsilon > 0$, such that F_x is twice differentiable on interval $B_\epsilon(\alpha, x) = (Q_\alpha(x) - \epsilon, Q_\alpha(x) + \epsilon)$ for every $x \in \Omega$ with first and second derivatives denoted as $f_x(\cdot)$ and $f_x'(\cdot)$ respectively. Furthermore, we assume $c_1 := \inf_{x\in\Omega, t\in B_\epsilon(\alpha,x)} f(x,t) > 0$, and $c_2 := \sup_{x\in\Omega, t\in B_\epsilon(\alpha,x)} |f'(x,t)| < \infty$.

(A3) There exist constants τ with $\lfloor \tau \rfloor > d/2$ and $c_3 > 0$ such that $\bar{\Phi}(w) \le c_3(1 + \|w\|^2)^{-\tau}$ for $w \in \mathbf{R}^d$, where $\bar{\Phi}$ is the Fourier transformation of Φ.

(A4) $c_4 m^{2\tau/d} \le n \le c_5 m^\gamma$ for constants $c_4, c_5 > 0$ and $\gamma \in (2\tau/d, \infty)$.

The next theorem formally states the asymptotic efficiency (proof is located in the supplementarily materials).

Theorem 1. Suppose (A1)–(A4) are met. If $\lambda_{m,n}^2 \sim (mn)^{-2\tau/(2\tau+d)}$ as $m, n \to \infty$, then $\|\hat{Q}_\alpha(\cdot) - Q_\alpha(\cdot)\|_{L^2(\Omega)} = O_p((mn)^{-\tau/(2\tau+d)})$.

Here, (A1) ensures the points in the design $\mathcal{X}$ will eventually fill the space as n grows, (A2) ensures the consistency and asymptotic normality of the sample quantile, (A3) is required to embed the reproducing kernel Hilbert space into a Sobolev space, and (A4) ensures a proper ratio of m and n to achieve efficiency.

The bound we establish agrees with the known optimal bounds (Stone 1982) for nonparametric regression, which implies that as $n, m \to \infty$, $a_{\lfloor \alpha m \rfloor}(x_0)$ approaches $Q_\alpha(x_0)$, and it does so at the fastest rate possible in terms of observed data. While previously developed techniques require the simulation output to be normally distributed, the efficiency shown in this section is *not* limited to the case when the simulation output is Gaussian.

Furthermore, this result addresses the question of replications in experiments for emulators. If the number of replications are properly related to the number of different inputs in the experiment, that is, $n \asymp m^\gamma$ where $\gamma > 2\tau/d$, we lose *no* efficiency in the emulator. Since τ is a measure of smoothness of the quantiles with respect to x, where large τ represents smooth quantiles, this result can be interpreted as: *if quantiles have little smoothness with respect to x, the experiment should consist of more replications.* This somewhat unintitive result is because the information gained by increasing n when studying a rough function is less than the information gained from replications.

The estimate in Equation (5) differs slightly from the one discussed in Section 3 because the covariance function is assumed to be specified. We present the following corollary, a direct result of Theorem 1, which explains the results in a more general context (similar ideas were presented in van der Vaart and van Zanten 2011).

Corollary 1. Suppose (A1)–(A4) hold and $Q_\alpha \in \mathcal{N}_\Phi(\Omega)$. Suppose that $\hat{Q}_\alpha^*$ is estimated by

$$\hat{Q}_\alpha^*(x) = \operatorname*{argmin}_{f \in \mathcal{N}_{\Phi*}(\Omega)} \frac{1}{n}\sum_{i=1}^{n}(\tilde{Q}_\alpha(x_i) - f(x_i))^2 + \lambda_{m,n}^2 \|f\|_{\mathcal{N}_{\Phi*}(\Omega)}^2,$$

where $\Phi^* \leq c_6\Phi$ for some $c_6 > 0$ and satisfies (A3) with a τ^*. If $\lambda_{m,n}^2 \sim (mn)^{-2\tau^*/(2\tau^*+d)}$ as $m, n \to \infty$, the following results hold:

- If $\tau^* = \tau$, then $\|\hat{Q}_\alpha^*(\cdot) - Q_\alpha(\cdot)\|_{L^2(\Omega)} = O_p((mn)^{-\tau/(2\tau+d)})$.
- If $\tau^* < \tau$ and $c_4 m^{2\tau^*/d} \leq n \leq c_5 m^\gamma$, then $\|\hat{Q}_\alpha^*(\cdot) - Q_\alpha(\cdot)\|_{L^2(\Omega)} = O_p((mn)^{-\tau^*/(2\tau^*+d)})$.

This demonstrates that even if the covariance function is misspecified, we can achieve the optimal convergence if we have correctly estimated the behavior of the covariance function, measured by τ. If we err on the conservative side and choose a covariance function with a small τ^*, for example, the exponential covariance function, we sacrifice efficiency for robustness. Importantly, the result for $\tau^* > \tau$ is not included above. Although it is not shown in Theorem 1, this condition could result in an inconsistent estimate.

5. ILLUSTRATIONS

Here, two examples are presented to illustrate the power of the proposed approach. The first deals with the crack propagation model discussed in Section 1. The second example provides a comparison between the proposed approach and the approach of Ankenman, Nelson, and Staum (2010) using the basic queueing system discussed in their work. Further details for implementation of the proposed method can be seen in the supplementary materials.

5.1 Material Fatigue

Fatigue of materials remains an important and challenging problem for engineers designing many structures from highways to turbine engines. Variability in loadings and material fatigue strength creates the need for a model that incorporates stochasticity. The study of fracture mechanics has recently focused on computational methods to understand propagation of faults in heterogeneous materials. Examples include piezoelectric materials (Li and Lee 2009) and complex composites (Grujicic et al. 2010); extensions to three-dimensional fractures have further increased the complexity of computational models (Jäger, Steinmann, and Kuhl 2008). The inclusion of stochasticity in these models makes computational techniques burdensome for fine mesh models. Here, we study the simplified case of one-dimensional crack propagation under transverse cyclic loading.

A classic deterministic model for the crack growth in this setting is the Forman equation (Forman, Kearney, and Engle 1967),

$$\frac{\mathrm{d}\ell}{\mathrm{d}t} = G(\ell) = f\frac{C_0(\Delta K(\ell))^n}{(1-R)K_c + \Delta K(\ell)},$$

where C_0 and n are constants, R is the ratio of the minimum ($S_{\min}$) and maximum ($S_{\max}$) pressures exerted on the material, f is the frequency of the cyclic loading, K_c is the fracture toughness, and $\Delta K(\ell) = (S_{\max} - S_{\min})\sqrt{\ell}\alpha(\ell)$. Here, $\alpha(\ell)$ is a function of the geometry and if the width of the structure is much larger than the crack size, $\alpha(\ell)$ can be approximated as 1. Though nonlinear, the Forman equation has the capability to represent both stable and accelerated growth rates (Stephens and Fuchs 2001). An extension of the deterministic model to account for variation is achieved multiplying the above growth rate by a stochastic process with unit mean (Lin and Yang 1983; Yang et al. 1983; Sobczyk 1986). Specifically, the model considered is

$$\mathrm{d}\ell = G(\ell)\mathrm{d}t + G(\ell)\mathrm{d}W(t),$$

where $W(t)$ is a realization of a Wiener process with variance σ^2. The material properties and conditions used in this simulation are borrowed from Hudson and Scardina (1969), which studied a plate of 7075 aluminum alloy with an initial crack length of 2.54 mm. This experiment will emulate the crack length after 2000 cycles under various stress ratios (further details can be seen in the supplementary materials).

Figure 2 shows emulators created with varying n and m. Subplot (a) shows an emulator with a small value of both n and m, which is not a good estimator of the true simulation seen in (d). Subplots (b) and (c) represent the improvements that occur as we increase n and m, respectively. From (a) to (b), n is increased and the shape of the quantiles as a function x is closer to the true shape of the quantiles shown in (d). From (b) to (c), m is increased allowing for better estimation of the individual quantiles (d) (compare the estimated quantiles at $R = 0$). For most inference, an emulator such as the one given in (c) will be sufficient for emulating the simulation.

5.2 Queueing System Example

We will now compare to Ankenman, Nelson, and Staum (2010) under their example, indicating the advantages of the proposed technique compared to the traditional metamodeling framework which assumes the simulation output follows a normal distribution. Here, we study a first-in first-out M/M/1 queue, that is a queue with one server and exponentially distributed interarrival (with mean x) and service times (with unit mean). The simulation output is the average system population in the system from time 0 to 1000. Ankenman, Nelson, and Staum (2010) defined the simulation output as a long-term average, but this work considers that this a finite horizon problem, which are commonly encountered (e.g., Ding, Puterman, and Bisi 2002). The experiment consists of n evenly spaced design points on [0.3, 0.9] with m replications. Figure 3 compares the emulative distribution of the proposed method to estimates established through dense sampling.

We quantitatively compare emulative distributions using *integrated quadratic distance* (IQD), which is a proper divergence score given by

$$\int_{-\infty}^{\infty}(F(y) - G(y))^2\mathrm{d}y,$$

where F is the emulative distribution and G is the actual distribution. Under some regulatory conditions (see Thorarinsdottir, Gneiting, and Gissibl 2013), IQD scores are equivalent to the metric

$$\mathbb{E}|R - S| - \frac{1}{2}\mathbb{E}|R - R'| - \frac{1}{2}\mathbb{E}|S - S'|,$$

where R and R' are independent copies from F, and S and S' are independent copies from G. A score of 0 indicates perfect emulation, and smaller values are preferred. Here, our goal is to develop an emulative distribution for a sample at a value x. Therefore, we create an average IQD (AIQD) by sampling

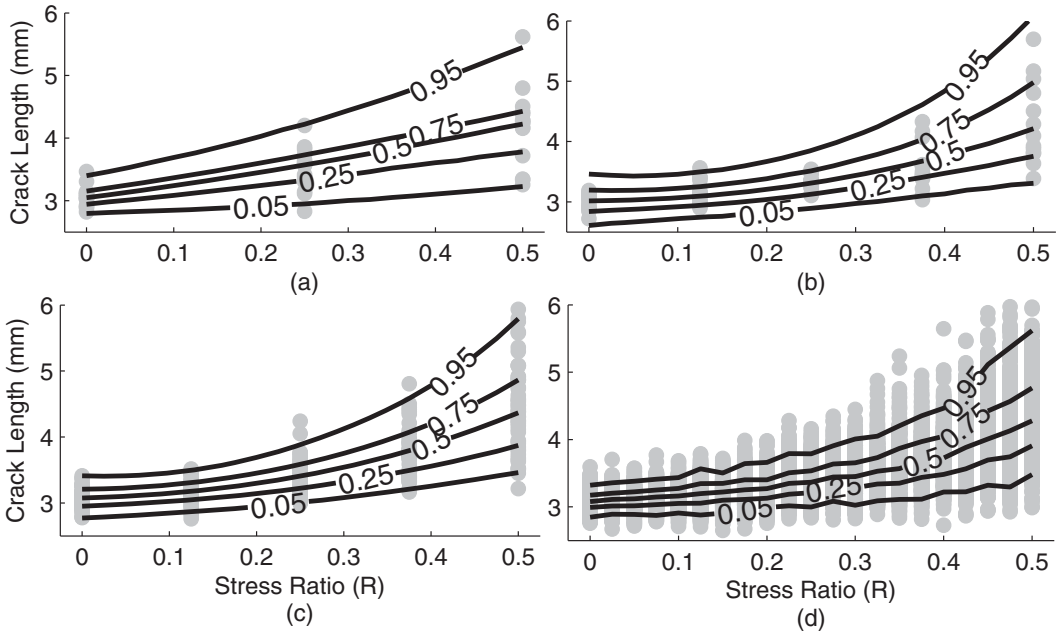

Figure 2. Example of emulation for Section 5.1; the light gray dots represent observations. Subplots (a), (b), and (c) contain the quantiles of the emulative distribution (solid line) with $n = 3$, $m = 15$ (a); $n = 5$, $m = 15$ (b); and $n = 5$, $m = 50$ (c). Subplot (d) represents empirical quantiles that are generated by simulating 400 observations at 20 points, requiring 8000 samples.

a value of x from $[0.3, 0.9]$. We create 400 replicates of the output at $[0.25, 0.275, \ldots, 0.925, 0.95]$ to create estimates of $\mathbb{E}|R - S|$ and $\mathbb{E}|S - S'|$.

Table 1 presents a comparison of AIQD using the proposed framework. The comparison is made using differing levels of m and n, and a smaller value represents superior prediction. Since the simulation is stochastic, it is difficult to compare values directly across m and n, though in general, increasing m and n results in better prediction. Clearly, the proposed method outperforms the traditional metamodeling framework, which is at least partially caused by instability in estimating $\sigma^2(x)$ as mentioned in Ankenman, Nelson, and Staum (2010).

6. DISCUSSION

Here, a framework is established for building emulators of stochastic simulations via *quantile kriging*, which enables a computationally attractive alternative to running the simulation model for every possible set of inputs. While emulators have been discussed to approximate specific models (e.g., Yang, Ankenman, and Nelson 2008), this work discusses constructions that do not rely on knowledge of the structure of the simulation. Other methods have been proposed to predict distributions, see De Iorio et al. (2004) and Dunson, Pillai, and Park (2007), but the focus in those works is on linear modeling, which can

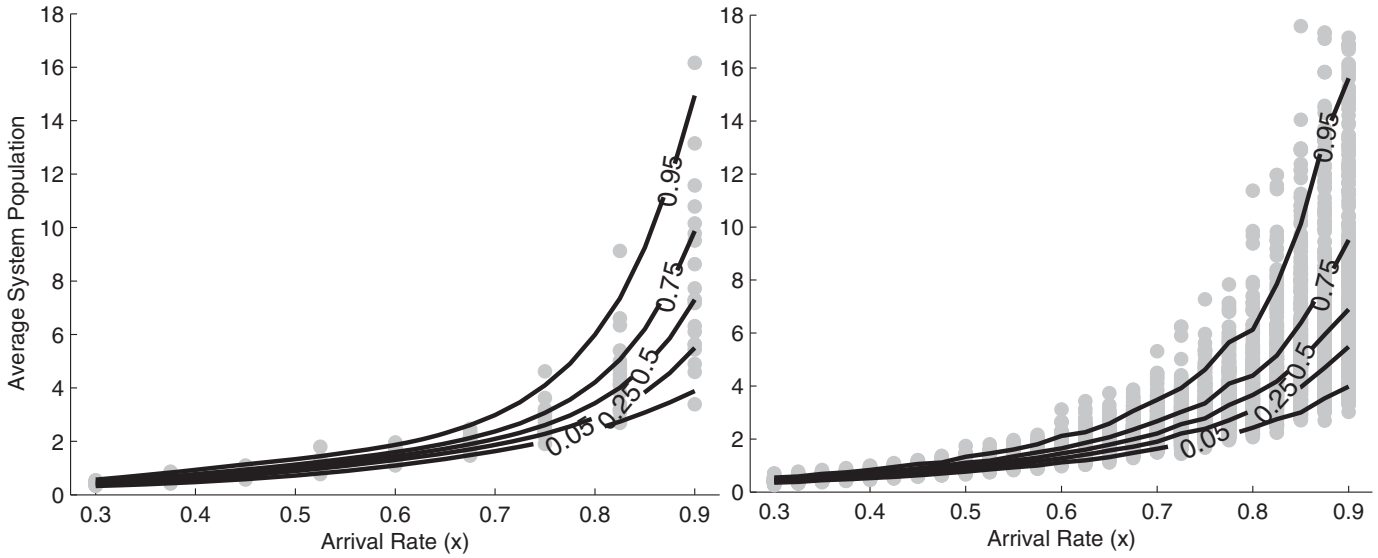

Figure 3. Example of an emulation for Section 5.2; the light gray dots represent observations. The left-hand plot contains the quantiles of the emulative distribution (solid line) with $n = 9$ and $m = 20$. The right-hand empirical quantiles are generated by simulating 400 observations at 27 points, requiring 10,600 samples.

Table 1. Value of AIQD (smaller is better) for the example in 5.2. "QK" represents the proposed method and "SK" represents a similar method proposed by Ankenman, Nelson, and Staum (2010)

n	m	QK	SK
5	10	0.0371	0.6270
	20	0.0342	0.7124
	40	0.0120	0.5875
9	10	0.0173	0.0796
	20	0.0192	0.1272
	40	0.0040	0.0939
17	10	0.0111	0.0653
	20	0.0064	0.0707
	40	0.0039	0.0681

exclude large classes of responses. A class of frameworks with a similar goal as us is known as quantile regression, which has been studied at great depth (for more information, see the text Koenker 2005). The focus of these techniques is adjustment of the loss function from a squared error loss to a piecewise linear loss to find estimates of individual quantiles. Of the works in quantile regression, Li, Liu, and Zhu (2007) is the closest to our work; the concept of their work is to add an additional penalty term representing the norm of a reproducing kernel Hilbert space, which is closely tied to Gaussian processes. This method incurs significant computational cost as the method requires quadratic programming to find each quantile, which can be burdensome for large amounts of data. This difficulty is exacerbated when finding unknown parameters, which requires hundreds or thousands of quadratic programs. Another relevant technique by Fan, Yao, and Tong (1996), often termed the "double kernel" approach, uses a similar two-stage mechanism as quantile kriging, but uses procedures involving locally defined polynomials. We yield to the comments of Wang and Wahba (1998) on the article Brumback and Rice (1998) who explain the difference in the modeling strategies; our work uses models that are defined over the entire region in lieu of locally defined models.

This work shows an asymptotic convergence rate of $O_p((nm)^{-\tau/(2\tau+d)})$, where τ is a measure of smoothness and d is the dimension of the input. This indicates that developing emulators will require a large sample size in high-dimensional scenarios, which means inversion of a large $n \times n$ matrix. Since inversion is the major obstacle in practice, we focus on its computational cost. For the proposed method, the number of arithmetic operations grows according to $O(n^3)$, which is an improvement over the methods such as Li, Liu, and Zhu (2007) which require $O(m^3n^3)$ operations when using scattered data. However, problems still arise if n increases. Works studying similar problems for deterministic computer codes, for example, Haaland and Qian (2011), might provide some insight into solutions.

More intricate choices of a set of inputs and replication balances for the designed experiment is outside the scope of this article, but challenges remain. In Section 4, we demonstrate an optimal rate of convergence by selecting the inputs via a uniform distribution and a symmetric number of replications. However, one would expect to achieve better small sample results using space-filling designs, such as those in Santner, Williams, and Notz (2003), to select sets of inputs for the experiment. Additionally, Ankenman, Nelson, and Staum (2010) emphasized the allocation of more replications to sets of inputs that produce high variation in the output. A similar approach might provide benefits here as well, but while this approach is justified when predicting the simulation output mean with normally distributed variations, the more general approach taken in this work adds complications.

ACKNOWLEDGMENTS

Plumlee's research is supported by the National Science Foundation grant CMMI-1030125. Tuo's research is supported by NSF grants DMS-0705261 and 1007574, and National Natural Science Foundation of China 11271355. The authors also thank C.F. Jeff Wu, V. Roshan Joesph, three anonymous reviewers, and the AE for their helpful comments on this work.

[Received February 2013. Revised October 2013.]

REFERENCES

Ankenman, B., Nelson, B. L., and Staum, J. (2010), "Stochastic Kriging for Simulation Metamodeling," *Operations Research*, 58, 371–382. [466,467,470,472]

Barton, R. R. (1998), "Simulation Metamodels," in *Simulation Conference Proceedings, 1998, Winter*, IEEE, Vol. 1, pp. 167–174. [466]

Brumback, B. A., and Rice, J. A. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves," *Journal of the American Statistical Association*, 93, 961–976. [472]

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953–963. [469]

De Iorio, M., Mueller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215. [471]

Diggle, P. J., and Ribeiro, P. J. (2007), *Model-Based Geostatistics* (Vol. 13), New York: Springer. [466]

Ding, X., Puterman, M. L., and Bisi, A. (2002), "The Censored Newsvendor and the Optimal Acquisition of Information," *Operations Research*, 50, 517–527. [470]

Dunson, D. B., Pillai, N., and Park, J. H. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society,* Series B, 69, 163–183. [471]

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206. [472]

Forman, R. G., Kearney, V. E., and Engle, R. M. (1967), "Numerical Analysis of Crack Propagation in Cyclic-Loaded Structures," *Journal of Basic Engineering*, 89, 459–463. [470]

Grujicic, M., He, T., Marvi, H., Cheeseman, B., and Yen, C. (2010), "A Comparative Investigation of the Use of Laminate-Level Meso-Scale and Fracture-Mechanics-Enriched Meso-Scale Composite-Material Models in Ballistic-Resistance Analyses," *Journal of Materials Science*, 45, 3136–3150. [470]

Haaland, B., and Qian, P. Z. (2011), "Accurate Emulators for Large-Scale Computer Experiments," *The Annals of Statistics*, 39, 2974–3002. [472]

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C., and Wilkinson, D. J. (2009), "Bayesian Emulation and Calibration of a Stochastic Computer Model of Mitochondrial DNA Deletions in Substantia Nigra Neurons," *Journal of the American Statistical Association*, 104, 76–87. [466]

Hudson, C. M., and Scardina, J. T. (1969), "Effect of Stress Ratio on Fatigue-Crack Growth in 7075-T6 Aluminum-Alloy Sheet," *Engineering Fracture Mechanics*, 1, 429–446. [470]

Jäger, P., Steinmann, P., and Kuhl, E. (2008), "On Local Tracking Algorithms for the Simulation of Three-Dimensional Discontinuities," *Computational Mechanics*, 42, 395–406. [470]

Kleijnen, J. P. C. (2007), *Design and Analysis of Simulation Experiments*, New York: Springer. [466,467]

Koenker, R. (2005), *Quantile Regression*, Cambridge: Cambridge University Press. [472]

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile Regression in Reproducing Kernel Hilbert Spaces," *Journal of the American Statistical Association*, 102, 255–268. [472]

Li, Y.-D., and Lee, K. Y. (2009), "Fracture Analysis on the Arc-Shaped Interface in a Layered Cylindrical Piezoelectric Sensor Polarized Along its Axis," *Engineering Fracture Mechanics*, 76, 2065–2073. [470]

Lin, Y., and Yang, J. (1983), "On Statistical Moments of Fatigue Crack Propagation," *Engineering Fracture Mechanics*, 18, 243–256. [470]

Matheron, G. (1963), "Principles of Geostatistics," *Economic Geology*, 58, 1246–1266. [466]

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013), "Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision," *Technometrics*, 55, 2–13. [466,467]

Ray, A., and Tangirala, S. (1996), "Stochastic Modeling of Fatigue Crack Dynamics for On-Line Failure Prognostics," *Control Systems Technology, IEEE Transactions*, 4, 443–451. [466]

Rigby, R., and Stasinopoulos, D. (2005), "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society,* Series C, 54, 507–554. [466]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–423. [466,467]

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer. [466,467,468,472]

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001), "A Generalized Representer Theorem," in *Computational Learning Theory*, Berlin: Springer, pp. 416–426. [469]

Sobczyk, K. (1986), "Modelling of Random Fatigue Crack Growth," *Engineering Fracture Mechanics*, 24, 609–623. [470]

Stephens, R. I., and Fuchs, H. O. (2001), *Metal Fatigue in Engineering*, New York: Wiley. [466,470]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [469]

Sze, D. Y. (1984), "OR Practice A Queueing Model for Telephone Operator Staffing," *Operations Research*, 32, 229–249. [466]

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N. (2013), "Using Proper Divergence Functions to Evaluate Climate Models" [online], available at *http://arxiv.org/abs/1301.5927*. [470]

van der Vaart, A., and van Zanten, H. (2011), "Information Rates of Nonparametric Gaussian Process Methods," *The Journal of Machine Learning Research*, 12, 2095–2119. [469]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59), Philadelphia, PA: Society for Industrial Mathematics. [469]

Wang, Y., and Wahba, G. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves: Comment," *Journal of the American Statistical Association*, 93, 976–980. [472]

Wendland, H. (2005), *Scattered Data Approximation*, Cambridge: Cambridge University Press. [469]

Yang, F., Ankenman, B. E., and Nelson, B. L. (2008), "Estimating Cycle Time Percentile Curves for Manufacturing Systems via Simulation," *INFORMS Journal on Computing*, 20, 628–643. [471]

Yang, J., Salivar, G., and Annis, C. (1983), "Statistical Modeling of Fatigue-Crack Growth in a Nickel-Base Superalloy," *Engineering Fracture Mechanics*, 18, 257–270. [470]