Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

ASYMMETRIC KRIGING EMULATOR FOR STOCHASTIC SIMULATION

Qiong Zhang

Department of Statistical Science and Operations Research
Virginia Commonwealth University
1015 Floyd Ave.
Richmond, VA 23284, USA

Wei Xie

Department of Industrial and System Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY 12180, USA

ABSTRACT

In many situations, e.g., simulation optimization and input uncertainty quantification, we need to assess the system performance at a large number of alternative inputs. Since each simulation run could be computationally expensive, statistical emulator could efficiently use the simulation budget to estimate the system performance. This paper proposes a new emulator for stochastic simulation, called asymmetric kriging (AK), which can be used to emulate the distribution of simulation outputs at each input point. Different from existing methods in the simulation literature, our approach does not require strong assumptions on either the functional form of the response surface or the normal distribution of the simulation estimation error. Numerical studies indicate the efficacy of our approach compared to alternative methods in the literature.

1 INTRODUCTION

Simulation is often used to assess the performance of stochastic systems characterized by various measures, e.g., mean, variance, probabilities, value-at-risk (VaR) and conditional value-at-risk (CVaR). Under many situations, we need to precisely estimate the system performance at a large number of alternative inputs. For example, given a pre-determined performance measure, simulation could be used to search for the optimal decision and the best system, e.g., finding the optimal ordering decision for an inventory system and selecting the best production system from hundreds candidates. Since each simulation run could be computationally expensive, given a potentially tight simulation resource, it is important to construct an emulator to quantify the input-output relationship. Compared to the direct simulation that runs the simulation at each candidate input to assess the system performance, the emulator constructed based on the simulation outputs at a few well-selected design points can aggregate the information on the response surface and further reduce the impact of simulation estimation uncertainty (Sun, Hong, and Hu 2014, Xie, Nelson, and Barton 2014). Thus, the emulator could efficiently employ the simulation resource to support optimization and input uncertainty quantification.

Parametric emulators can be used when we have strong prior information on the parametric form of system response surface or it is built as a local metamodel based on the Taylor series approximation; see

the review in Henderson and Nelson (2006). However, in many situations, we do not have strong prior information on the response surface. Stochastic kriging (SK) introduced in Ankenman, Nelson, and Staum (2010) can overcome the limitations of parametric emulators, and it does not require any strong prior information of the system response surface. However, SK relies on the assumption that the simulation estimation error follows the normal distribution, which does not hold in many situations. For example, when we study the expected number of customers in the system for an $M/M/1$ queue, this normal assumption is not appropriate when the utilization is high and the runlength is short. For a risk measure, e.g., the 99%th percentile of first 100 customers' times staying in the queue, it could be hard to meet the normal assumption.

The important properties in the simulation estimation uncertainty, such as skewness and tails, could impact on the sampling distribution when the emulator of system performance is used to guide the search for the optimal decision (Sun, Hong, and Hu 2014). They could also impact the percentile credible interval quantifying both input and simulation estimation uncertainty (Xie, Nelson, and Barton 2014). Thus, it is necessary to construct a distributional emulator for the simulation outputs to capture these important properties in simulation estimation uncertainty.

Without assuming any parametric form on the distribution of simulation estimation error, Plumlee and Tuo (2014) developed quantile Kriging (QK) to construct a distributional emulator for the simulation output. At each input point, QK provides an empirical distribution of stochastic simulation outputs based on the fitted quantile curves over the input space. These quantile curves are obtained by interpolating corresponding empirical quantiles at all design points. Therefore, the performance of QK greatly depends on the accuracy of the empirical quantile estimation, which typically requires a large number of replications at each design point.

In this paper, we develop a new method, called Asymmetric Kriging (AK), to construct the distributional emulator. It does not require any strong prior information on the response surface and also the normal assumption on the simulation estimation error. Our study indicates that AK provides a good performance even when a large number of replications is not affordable. Specifically, similar to QK, the distributional emulator of AK is built on multiple quantile curves over the entire design space. The quantile curves are obtained by solving the functional based asymmetric least squares under different asymmetric weights (Newey and Powell 1987). Each quantile curve is fitted by coupling all simulation outputs together. Different from QK, AK does not rely on the empirical quantiles at each design point. Thus, we do not require a large number of replications at each design point to guarentee the accuracy of the AK based distributional emulator. The empirical study in Section 5 demonstrates that AK has promising finite-sample performance.

The next section provides a formal problem description. We use an $M/M/1$ queue as a motivating example to illustrate the limitation of the normal assumption on the simulation estimation error. Section 3 reviews QK, and Section 4 introduces our distributional emulator. The same $M/M/1$ queue system is used to study the finite-sample performance of SK, QK and our approach in Section 5. We give concluding remarks in Section 6. Proof for our theoretical result is deferred to the Appendix.

2 PROBLEM DESCRIPTION AND A MOTIVATING EXAMPLE

We consider stochastic simulation models with a continuous scalar output. At each input point $\mathbf{x}$ in the input space $\mathscr{X}$, the simulation output can be modeled by

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

where the input point $\mathbf{x}$ is a vector, including the decision variables and the estimates of input models, e.g., the rates of the inter-arrival and service time distributions in an $M/M/1$ queue, $\mu(\mathbf{x})$ represents the unknown system performance measure of interest, and $\varepsilon(\mathbf{x})$ is a zero-mean random variable representing the simulation estimation error. *Our goal is to emulate the distribution of* $Y(\mathbf{x})$ *at each input point* $\mathbf{x} \in \mathscr{X}$ *so that we can eventually capture important properties in the estimation uncertainty of the response* $\mu(\mathbf{x})$.

SK can be used to emulate the unknown response surface $\mu(\mathbf{x})$. Under this framework, $\mu(\mathbf{x})$ has a Gaussian process (GP) prior with mean μ_0 and covariance structure $\sigma^2 c(\mathbf{x},\mathbf{x}')$, where σ^2 denotes the variance and $c(\cdot,\cdot)$ denotes a correlation function. Given the design points $\mathscr{D} \equiv \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k\} \subset \mathscr{X}$, we obtain simulation data

$$\mathscr{T} = \{\mathbf{x}_i, [Y_1(\mathbf{x}_i), Y_2(\mathbf{x}_i), \ldots, Y_{n_i}(\mathbf{x}_i)]\}_{i=1}^k, \tag{2}$$

where n_i denotes the number of replications at design point $\mathbf{x}_i$. Let $\bar{\mathbf{Y}}_{\mathscr{D}} = [\bar{Y}(\mathbf{x}_1), \bar{Y}(\mathbf{x}_2), \ldots, \bar{Y}(\mathbf{x}_k)]^\top$ with $\bar{Y}(\mathbf{x}_i) = \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i)/n_i$. Denote the variance of $\bar{\mathbf{Y}}_{\mathscr{D}}$ as a $k \times k$ diagonal matrix C with i-th diagonal entry $\sigma_\varepsilon^2(\mathbf{x}_i)/n_i$. Then, the uncertainty of the response surface $\mu(\mathbf{x})$ can be quantified by the posterior normal distribution of $\mu(\mathbf{x})$ with mean

$$m_p(\mathbf{x}) = \widehat{\mu}_0 + \mathbf{r}_k^\top(\mathbf{x})(\Sigma_k + \sigma^{-2}C)^{-1}(\bar{\mathbf{Y}}_{\mathscr{D}} - \widehat{\beta}_0 \mathbf{1}_k), \tag{3}$$

and variance

$$\sigma_p^2(\mathbf{x}) = \sigma^2 - \sigma^2 \mathbf{r}_k^\top(\mathbf{x})(\Sigma_k + \sigma^{-2}C)^{-1}\mathbf{r}_k(\mathbf{x}) + \eta^\top [1_{k\times 1}^\top (\sigma^2\Sigma_k + C)^{-1}\mathbf{1}_k]^{-1}\eta,$$

where Σ_k is the $k \times k$ correlation matrix of the design points, $\mathbf{r}_k(\mathbf{x})$ is the $k \times 1$ correlation vector between each design point and a fixed prediction point $\mathbf{x}$, $\widehat{\mu}_0 = [\mathbf{1}_k^\top(\Sigma_k + \sigma^{-2}C)^{-1}\mathbf{1}_k]^{-1}\mathbf{1}_k^\top(\Sigma_k + \sigma^{-2}C)^{-1}\bar{\mathbf{Y}}_{\mathscr{D}}$ and $\eta = 1 - \mathbf{1}_k^\top(\Sigma_k + \sigma^{-2}C)^{-1}\mathbf{r}_k(\mathbf{x})$.

In the SK metamodel, the normal assumption of the stochastic simulation errors is often required, $\varepsilon(\mathbf{x}) \sim N(0, \sigma_\varepsilon^2(\mathbf{x}))$, which does not hold in general. Here, we use an $M/M/1$ example to illustrate that there often exist skewness and tails in the simulation estimation error.

Example: We study the performance of the first $M = 300$ customers' times staying in an $M/M/1$ queue system. The system starts with empty. We fix the arrival rate $\mu_A = 1$, and vary the service rate from 1.1 to 3.5. The input x is the mean service time. Let $Q_1, Q_2, \ldots, Q_M$ denote the times of the first M customers staying in the system. From each replication, the sample mean and the 99%-th sample quantile of $Q_1, Q_2, \ldots, Q_M$ are used to estimate the mean and the 99% VaR. In Figure 1, we show the qqnorm plots obtained from 400 replications of simulation outputs at $x = 0.5$. Based on our empirical study, we observe stronger skewness and tails as the utilization becomes closer to one.

Given a tight simulation budget, these important properties, including skewness and tails, could impact the estimation of the response $\mu(\mathbf{x})$. Thus, in this paper, we introduce a new distributional emulator that does not require strong prior information on the response surface $\mu(\mathbf{x})$ and also the distribution family of simulation estimation uncertainty.

3 QUANTILE KRIGING BASED DISTRIBUTIONAL EMULATOR

Quantile Kriging (Plumlee and Tuo 2014) provides a convenient way to construct distributional emulator for a stochastic simulation system. The basic idea is to build a distributional emulator using the sample quantiles at each design point. This section will review the methodology and some computational details of QK. We refer to Plumlee and Tuo (2014) for the theoretical justification of QK.

For notational convenience, Plumlee and Tuo (2014) assume that the number of replications at each design point are all equal, say, $n = n_1 = \ldots = n_k$ in $\mathscr{T}$ defined by (2). At each design point $\mathbf{x}_i$ in $\mathscr{D}$, let $Y_{(j)}(\mathbf{x}_i)$ be the jth order statistic of the n replications. Also, $Y_{(j)}(\mathbf{x}_i)$ is the α-th sample quantile with $\alpha \in [(j-1)/n, j/n)$. Denote $\mathbf{Y}_{(j)}$ as a vector containing the j-th order statistics at all k design points:

$$\mathbf{Y}_{(j)} = \left[Y_{(j)}(\mathbf{x}_1), \ldots, Y_{(j)}(\mathbf{x}_k)\right]^\top.$$

The estimated α-th percentile curve would be the simple Kriging predictor (Cressie 2015) fitted using $\mathbf{Y}_{(j)}$

$$\hat{Y}_j(\mathbf{x}) = \mu_0 + \mathbf{r}_k(\mathbf{x})^\top(\Sigma_k + \rho I_k)^{-1}(\mathbf{Y}_{(j)} - \mu_0) \text{ for } \mathbf{x} \in \mathscr{X}. \tag{4}$$

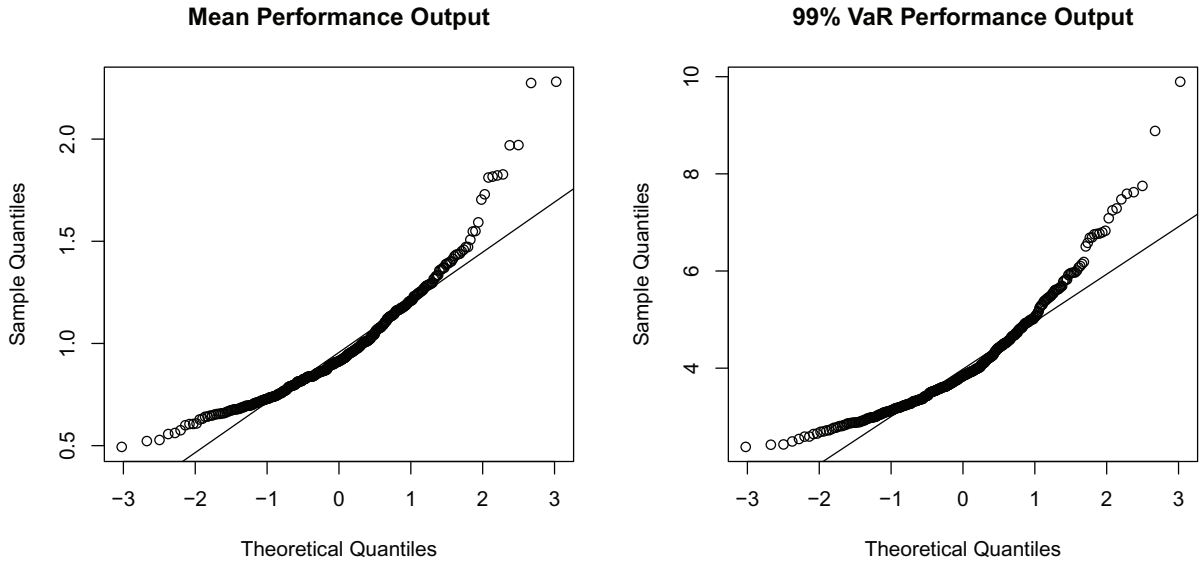

Figure 1: The qqnorm plots of the mean performance and 99%-th VaR performance outputs at $x = 0.5$ for the $M/M/1$ queue.

As noted earlier, μ_0 is the constant mean of the simulation system, $\mathbf{r}_k(\mathbf{x})$ is a $k \times 1$ correlation vector and Σ_k is a $k \times k$ correlation matrix. Given a correlation function $c(\cdot,\cdot)$, the j-th entry in $\mathbf{r}_k(\mathbf{x})$ is $c(\mathbf{x},\mathbf{x}_i)$, and the (i,i')-th entry in Σ_k is $c(\mathbf{x}_i,\mathbf{x}_{i'})$. The unknown parameter $\rho(>0)$ is the nugget parameter, and I_k is the $k \times k$ identity matrix.

After fitting the quantile curves for $j = 1,\ldots,n$, the unknown parameters in (4), such as the nugget parameter ρ and the parameters in the correlation function $c(\cdot,\cdot)$, can be estimated using the leave-one-out cross-validation. In the numerical implementation of Plumlee and Tuo (2014), the unknown parameters are assumed to be the same for the Kriging predictors of all the quantile curves, and the optimal parameters are selected as the minimizer of the sum of the leave-one-out cross-validation errors over all the quantile Kriging predictors; See Section 3.3 of Plumlee and Tuo (2014) for more detail. The Kriging predictor $\hat{Y}_j(\mathbf{x})$ in (4) for $j = 1,\ldots,n$ forms a distributional emulator. Specifically, for any $\mathbf{x} \in \mathscr{X}$, we obtain an empirical cumulative distribution function

$$\widehat{F}_{\mathbf{x}}(t) = \frac{1}{n}\sum_{j=1}^{n} I\left[t \le \hat{Y}_j(\mathbf{x})\right],$$

where $I(\cdot)$ is the indicator function. We further define

$$\widehat{\mathbf{F}}(\cdot) = \left\{\widehat{F}_{\mathbf{x}}(\cdot)|\mathbf{x} \in \mathscr{X}\right\} \tag{5}$$

as the distributional emulator based on QK. The estimates for mean or quantiles at each input point would be derived based on this distribution emulator. For example, the mean response surface can be expressed by

$$\hat{\mu}(\mathbf{x}) = \int Y d\widehat{F}_{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \hat{Y}_j(\mathbf{x}) = \mu_0 + \mathbf{r}_k(\mathbf{x})^\top (\Sigma_k + \rho I_k)^{-1}\left[\bar{\mathbf{Y}}_{\mathscr{D}} - \mu_0\right],$$

and the $\alpha \in [(j-1)/n, j/n)$th quantile estimates is $\hat{Y}_j(\mathbf{x})$ as in (4).

QK is developed based on the order statistics at each design point. If the number of replications is small, the order statistics do not accurately estimate the quantiles. The estimation can be extremely inaccurate when we are interested in fitting a large number of quantile curves with a limited number of replications. To solve this challenge, we propose a new method to the distributional emulator in Section 4.

4 ASYMMETRIC KRIGING BASED DISTRIBUTIONAL EMULATOR

4.1 Quantile and Expectile

Motivated by QK, the key of constructing a distributional emulator is to fit a large number of quantile curves of simulation output Y over the design space $\mathscr{X}$. In the literature of statistics, both quantile regression (Koenker 2005) and asymmetric least squares based expectile regression (Newey and Powell 1987) can be used to obtain quantile curves. The main difference of these two methods is their loss functions used to quantify the distance between fitted response f and the data Y. The loss function used in quantile regression is

$$Q_\tau(Y,f) = |Y-f|\{\tau I[Y \le f] + (1-\tau) I[Y > f]\},$$

and the loss function used in asymmetric least squares is

$$Q_\tau(Y,f) = (Y-f)^2\{\tau I[Y \le f] + (1-\tau) I[Y > f]\}, \tag{6}$$

where $0 < \tau < 1$ is a weight parameter. According to (6), asymmetric least square modifies the traditional least squared based regression by adding asymmetric weights. By solving $\min_f \mathrm{E} Q_\tau(Y,f)$, the fitted model f can be expressed by

$$\hat{f}_\tau = \frac{\tau \mathrm{E}[YI(Y \le f)] + (1+\tau)\mathrm{E}[YI(Y > f)]}{\tau P(Y \le f) + (1-\tau)P(Y > f)},$$

where the expectation and probability are taken with regard to the distribution of Y. Due to the availability of this explicit form, expectile regression has become a popular approach in risk management (Kuan, Yeh, and Hsu 2009).

Asymmetric least squares has been used to compute multiple quantile curves in analyzing simple observational data; see examples in Efron (1991). Although the linear or polynomial model in Efron (1991) can be used for observation data, these models might be too simple to describe the complex input-output relationship in simulation analysis. Hence, we propose asymmetric Kriging to address this issue by including the kernel bases functions into the asymmetric least squares regression.

4.2 Kernel Correlation Function Based Regression

Consider the dataset $\mathscr{T}$ in (2). Let $N = \sum_{i=1}^k n_i$. Denote $\mathbf{Y} = (Y^1, \ldots, Y^N)$ as a vector containing all the outputs. For $i = 1, \ldots, N$, we define $\mathbf{x}^i$ as the input point corresponding to the output Y^i. The predictor of regularized least squares regression is obtained by minimizing

$$L(f) = \sum_{i=1}^N \left[Y^i - f(\mathbf{x}^i)\right]^2 + \rho \langle f, f \rangle \text{ for } f \in \mathscr{F}, \tag{7}$$

where ρ is a tuning parameter, the functional space $\mathscr{F}$ is defined by

$$\mathscr{F} = \left\{ f \in R^{\mathscr{X}} \,\middle|\, f(\cdot) = \sum_{i=1}^\infty \beta_i c(\cdot, z_i), \beta_i \in R, z_i \in \mathscr{X}, \langle f, f \rangle < \infty \right\}, \tag{8}$$

with $c(\mathbf{x}^i, \mathbf{x}^j)$ be a correlation function defined in (4), and the inner product $\langle f, g \rangle$ is defined by

$$\langle f, g \rangle := \sum_{i=1}^m \sum_{j=1}^{m'} \beta_i \gamma_j c(\mathbf{z}_i, \mathbf{z}_j)$$

for two arbitrary components $f(\cdot) = \sum_{i=1}^{m} \beta_i c(\cdot, \mathbf{z}_i)$ and $g(\cdot) = \sum_{j=1}^{m'} \gamma_j c(\cdot, \mathbf{z}_j)$ in $\mathscr{F}$.

According to the representer theorem (e.g., Wahba (1990) and Schölkopf, Herbrich, and Smola (2001)), the solution of minimizing the objective function in (7) admits a representation

$$f(\mathbf{x}) = \sum_{i=1}^{N} \beta_i c(\mathbf{x}, \mathbf{x}^i) = \boldsymbol{\beta}^\top \mathbf{r}_N(\mathbf{x}).$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^\top$, and $\mathbf{r}_N(\mathbf{x})$ is an N-dimensional vector with i-th component $c(\mathbf{x}, \mathbf{x}^i)$. Therefore, $L(f)$ in (7) can be expressed by

$$L(\boldsymbol{\beta}) = (\mathbf{Y} - \Sigma_N \boldsymbol{\beta})^\top (\mathbf{Y} - \Sigma_N \boldsymbol{\beta}) + \rho \boldsymbol{\beta}^\top \Sigma_N \boldsymbol{\beta},$$

where Σ_N is an $N \times N$ matrix with the (i, j)th component $c(\mathbf{x}^i, \mathbf{x}^j)$. Minimizing $L(\boldsymbol{\beta})$ with regard to $\boldsymbol{\beta}$, we obtain

$$\hat{\boldsymbol{\beta}} = (\Sigma_N + \rho I_N)^{-1} \mathbf{Y},$$

where I_N is the $N \times N$ identify matrix. The predictor developed from (7) is

$$\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\beta}}^\top \mathbf{r}_N(\mathbf{x}) = \mathbf{r}_N(\mathbf{x})^\top (\Sigma_N + \rho I_N)^{-1} \mathbf{Y}. \tag{9}$$

Computing $\hat{f}(\mathbf{x})$ requires solving a linear system of size N. We consider how to reduce this calculation when there are replications in the data. Let $\mathbf{Y}_i$ be a vector of size n_i collecting all the outputs from input point $\mathbf{x}_i$. We can alternatively express $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_k^\top)^\top$. Following the notation in Binois, Gramacy, and Ludkovski (2016), Σ_N can be expressed by

$$\Sigma_N = U \Sigma_k U^\top, \tag{10}$$

where Σ_k is a $k \times k$ covariance matrix constructed by the k unique input points, and U is an $N \times k$ sampling matrix:

$$U = \text{diag}\{\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_k}\},$$

with $\mathbf{1}_{n_i}$ being a vector of size n_i loaded by ones. Similarly, we can represent $\mathbf{r}_N(\mathbf{x})$ as $U\mathbf{r}_k(\mathbf{x})$. According to the Woodbury identity and the development in Binois, Gramacy, and Ludkovski (2016), $\hat{f}(\mathbf{x})$ in (9) is reduced to

$$\hat{f}(\mathbf{x}) = \mathbf{r}_k(\mathbf{x})^\top \left[\Sigma_k + \rho \text{diag}(1/n_1, \ldots, 1/n_k)\right]^{-1} \bar{\mathbf{Y}}_{\mathscr{D}}, \tag{11}$$

It is easy to see that $\hat{f}(\mathbf{x})$ is equivalent to the posterior mean of $\mu(\mathbf{x})$ in (3) when μ_0 is zero and $\sigma_\varepsilon^2(\mathbf{x}) = \rho$ for all $\mathbf{x} \in \mathscr{X}$.

The predictor in (9) only provides the mean estimates of the output at each input point. However, a distributional emulator is constructed using multiple predictors each representing a quantile curve. The idea of asymmetric Kriging is to replace the least squares loss in (7) with the asymmetric loss. Multiple predictors could be fitted by varying the value of τ in (6). We now propose SK based on this idea.

4.3 Our Proposal: Asymmetric Kriging

We generalize (7) using the asymmetric least squares loss function in (6):

$$L_\tau(f) = \sum_{i=1}^{N} Q_\tau(Y^i, f(\mathbf{x}^i)) + \rho \langle f, f \rangle \text{ for } f \in \mathscr{F}, \tag{12}$$

where $\mathscr{F}$ is the same functional space in (8). By the representer theorem, the solution of (12) also can be expressed by

$$f(\mathbf{x}) = \boldsymbol{\beta}_\tau^\top \mathbf{r}_N(\mathbf{x}).$$

Therefore, minimizing (12) with regard to f is equivalent to minimizing

$$L(\boldsymbol{\beta}_\tau) = (\mathbf{Y} - \Sigma_N \boldsymbol{\beta}_\tau)^\top W_\tau (\mathbf{Y} - \Sigma_N \boldsymbol{\beta}_\tau) + \boldsymbol{\beta}_\tau^\top \Sigma_N \boldsymbol{\beta}_\tau$$

with regard to $\boldsymbol{\beta}_\tau$. The weight matrix W_τ above is an N by N diagonal matrix with the i-th diagonal entry

$$w_i = \tau I[Y^i \leq f(\mathbf{x}^i)] + (1-\tau) I[Y^i > f(\mathbf{x}^i)].$$

By taking derivative with regard to $\boldsymbol{\beta}_\tau$, the solution of minimizing $L(\boldsymbol{\beta}_\tau)$ is

$$\hat{\boldsymbol{\beta}}_\tau = (\Sigma_N + \rho W_\tau^{-1})^{-1} \mathbf{Y}, \tag{13}$$

which further leads to

$$\hat{f}_\tau(\mathbf{x}) = \mathbf{r}_N(\mathbf{x})^\top (\Sigma_N + \rho W_\tau^{-1})^{-1} \mathbf{Y}. \tag{14}$$

Since the predictor $\hat{f}_\tau(\mathbf{x})$ combines the Kriging and the asymmetric least squares regression, we call it asymmetric Kriging. The mean estimates at each design point is $\hat{f}_\tau(\mathbf{x})$ with $\tau = 0.5$, which can be obtained directly without iterative computing. The percentile of $\hat{f}_\tau(\mathbf{x})$ can be estimated by

$$\hat{\alpha}_\tau = \frac{1}{N} \sum_{i=1}^{N} I[Y^i \leq \hat{f}_\tau(\mathbf{x}^i)].$$

Therefore, $\hat{f}_\tau(\mathbf{x})$ is the estimated $\hat{\alpha}_\tau$-th quantile curve. By combining multiple quantile curves with different percentiles, we obtain a distributional emulator as in (5).

Similar as (11), we can reduce the calculation of $N \times N$ matrix in (14) for simulation outputs with replications. The reduced emulator is presented in Proposition 1.

Proposition 1 Let Λ_τ be a diagonal matrix of size k (number of design points) with the i-th diagonal entry

$$\lambda_i = \tau \sum_{j=1}^{n_i} I\left[Y_j(\mathbf{x}_i) \leq f_\tau(\mathbf{x}_i)\right] + (1-\tau) \sum_{j=1}^{n_i} I\left[Y_j(\mathbf{x}_i) > f_\tau(\mathbf{x}_i)\right].$$

Let $\bar{\mathbf{Y}}_\tau$ be a vector of size k with i-th entry

$$\bar{Y}_{i,\tau} = \lambda_i^{-1} \left\{ \tau \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i) I\left[Y_j(\mathbf{x}_i) \leq f_\tau(\mathbf{x}_i)\right] + (1-\tau) \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i) I\left[Y_j(\mathbf{x}_i) > f_\tau(\mathbf{x}_i)\right] \right\}.$$

Then the asymmetric kriging predictor in (14) can be reduced to

$$\hat{f}_\tau(\mathbf{x}) = \mathbf{r}_k(\mathbf{x})^\top (\Sigma_k + \rho \Lambda_\tau^{-1})^{-1} \bar{\mathbf{Y}}_\tau. \tag{15}$$

The proof of Proposition 1 is deferred to the Appendix. Since both $\bar{\mathbf{Y}}_\tau$ and Λ_τ depend on $f_\tau(\mathbf{x})$, we can obtain $\hat{f}_\tau(\mathbf{x})$ by computing $\mathbf{r}_k(\mathbf{x})^\top (\Sigma_k + \rho \Lambda_\tau^{-1})^{-1} \bar{\mathbf{Y}}_\tau$ iteratively till convergence. As in Plumlee and Tuo (2014), we also use the leave-one-out cross-validation method to determine the optimal values of unknown parameters ρ and correlation parameters in the kernel correlation function $c(\cdot,\cdot)$. In our numerical implementation, we assume that these unknown parameters are the same across different percentile curves. Therefore, their optimal values are obtained by minimizing the sum of leave-one-out cross-validation errors over all the percentile curves.

5 EMPIRICAL STUDY

We revisit the $M/M/1$ queue system example in Section 2, and compare three methods:

- **SK:** Stochastic Kriging in Section 2,
- **QK:** Quantile Kriging in Section 3,
- **AK:** Asymmetric Kriging in Section 4.

For all three methods, we use the Gaussian correlation function to quantify the dependence between input points. We consider the mean, the 95%-th and the 99%-th VaR as examples of simulation outputs. In all our numerical experiments, we generate the input design as equally spaced points from 0.3 to 0.9.

We first graphically present the distributional emulators constructed by all three methods. Five design points each with five replications are generated as the input points. We generate simulation outputs as the mean performance, the 95%-th VaR performance, and 99%-th VaR performance. The 0.1th, 0.25th, 0.5th, 0.75th and 0.95th quantile curves are obtained based on the fitted distributional emulators as shown in Figure 2. We see that, the AK fitted quantile curves are more smooth compared to SK and QK. Also, due to the insufficient number of replications, the quantile curves from QK exhibit crossing or overlapping many times.

We now investigate the overall performances of the distributional emulator. To compare the fitted distributions and the actual distributions, we use a large test dataset to empirically estimate the actual distributions at 100 equally spaced input points. At each of these input points, we create 400 replicates of the outputs to ensure the accuracy of the empirical distributions. We first quantitatively measure the distance between the emulated distribution and the true distribution provided by the test data sets. According to Plumlee and Tuo (2014), integrated quadratic distance (IQD) can be used to measure the distance of two distributions. Let $F(v)$ and $G(v)$ be the cumulative density functions of two distributions, IQD is given by

$$\int_{-\infty}^{\infty} \{F(v) - G(v)\}^2 dv,$$

Following Thorarinsdottir, Gneiting, and Gissibl (2013), IQD can be alternatively expressed by

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

where R and R' are independent samples from $F(v)$, and S and S' are independent samples from $G(v)$. At each input point in the test dataset, IQD can be estimated using the independent samples generated from the true distribution and the emulated distribution. We then compute the average IQD (AIQD) by taking average of the IQD values over all input points in the test dataset. The resulted AIQD measures the average distance between the emulated distribution and the true distribution over the entire input space. All numerical experiments are repeated for 100 times. In Table 1, we show the average results of AIQD for each method under different settings. We see that, the smallest AIQDs consistently come from AK (the proposed method).

We also measure the overall accuracy of estimated quantile curves. Based on the test datasets, the empirical quantiles can be estimated at each design point. For each method, the average mean squared error (AMSE) over all quantile curves and all design points in the test datasets can be calculated by

$$\mathrm{AMSE} = \frac{1}{100 \times 99} \sum_{i=1}^{100} \sum_{j=1}^{99} (\hat{q}_{ij} - \tilde{q}_{ij})^2,$$

where $\hat{q}_{ij}$ is the j%-th empirical quantile at the i-th input point in the test data set, and $\tilde{q}_{ij}$ is the corresponding emulated quantile obtained by one of the three methods. AMSE provides the overall performance of the distributional emulator in estimating quantile curves. In Table 2, we show the average results of AMSE

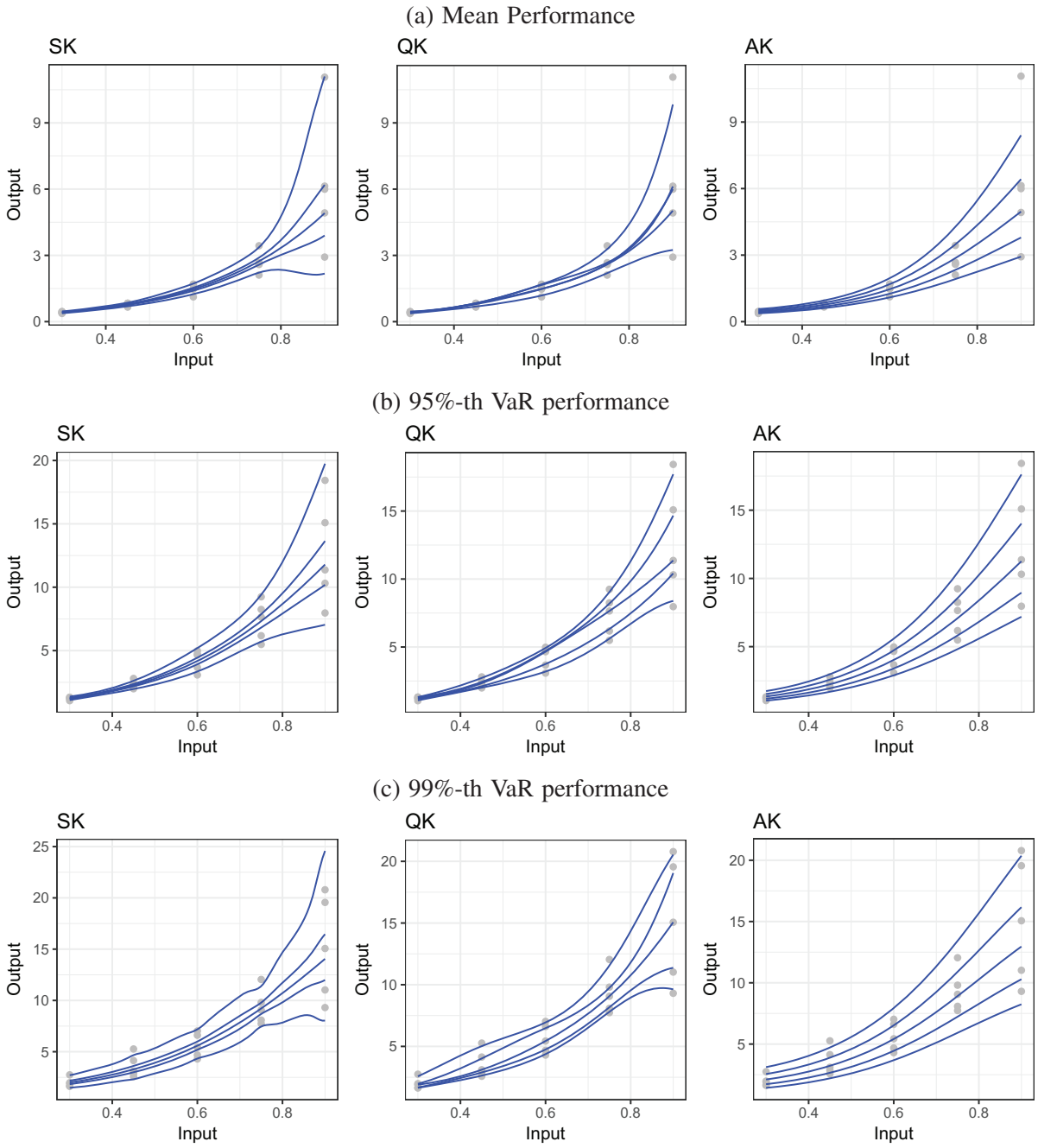

Figure 2: The 0.1th, 0.25th, 0.5th, 0.75th and 0.95th quantile curves provided by the three methods for each case (Note: grey dots are the simulation data used to fit these quantile curves).

for each method under different settings over 100 macro-replications. Similar to the AIQD results, AK achieves the best performance among all three methods. Compared to QK, the advantage of AK is more significant when the number of replication at each design point is small.

Table 1: The AIQD results calculated for each method under different settings over 100 macro-replications.

Case	#Design	# replication=5			# replication=10		
		AK	QK	SK	AK	QK	SK
Mean	5	**0.494**	0.537	0.539	**0.485**	0.513	0.540
	9	**0.486**	0.516	0.536	**0.488**	0.506	0.542
95% VaR	5	**1.127**	1.163	1.188	**1.093**	1.114	1.168
	9	**1.103**	1.133	1.200	**1.076**	1.085	1.167
99% VaR	5	**1.342**	1.385	1.425	**1.303**	1.332	1.407
	9	**1.319**	1.356	1.434	**1.283**	1.298	1.404

Table 2: The AMSE results calculated for each method under different settings over 100 macro-replications.

Case	#Design	# replication=5			# replication=10		
		AK	QK	SK	AK	QK	SK
Mean	5	**0.517**	0.781	0.692	**0.368**	0.508	0.548
	9	**0.447**	0.645	0.652	**0.371**	0.461	0.591
95% VaR	5	**2.057**	2.525	2.659	**0.541**	1.681	2.341
	9	**1.844**	2.331	2.924	**1.209**	1.216	2.337
99% VaR	5	**2.221**	2.804	3.040	**1.614**	1.868	2.879
	9	**1.960**	2.686	3.407	**1.253**	1.345	2.862

6 CONCLUDING REMARKS

We propose a new method, asymmetric Kriging (AK), to construct the distributional emulator for simulation analysis. AK fits multiple quantile curves to the simulation data, and returns a distributional emulator. It does not require that the simulation estimation error follows a normal distribution. Mean and quantile curves can be efficiently estimated using this method. Future directions of our proposal are indicated as follows: 1) develop methods to quantify the uncertainty of the distributional emulator under the Bayesian framework; 2) investigate the theoretical properties of the proposed method; 3) apply our method to complex simulation systems, such as supply chains, manufacturing and service systems.

A PROOF OF PROPOSITION 1

Following the notation in Binois, Gramacy, and Ludkovski (2016), let U be the $N \times k$ sampling matrix in (10). The asymmetric kriging predictor in (14) can be expressed by

$$\hat{f}_\tau(\mathbf{x}) = \mathbf{r}_k^\top(\mathbf{x}) U^\top (U\Sigma_k U^\top + \rho W_\tau^{-1})^{-1} \mathbf{Y},$$

where $\mathbf{r}_k(\mathbf{x})$ and Σ_k are defined in (3). By using the Woodbury identity to inverse $U\Sigma_k U^\top + \rho W_\tau^{-1}$, we obtain

$$\hat{f}_\tau(\mathbf{x}) = \mathbf{r}_k^\top(\mathbf{x}) U^\top \left[\rho^{-1} W_\tau - \rho^{-1} W_\tau U (\rho\Sigma_k^{-1} + U^\top W_\tau U)^{-1} U^\top W_\tau\right] \mathbf{Y},$$

which can be further expressed as

$$\hat{f}_\tau(\mathbf{x}) = \rho^{-1}\mathbf{r}_k^\top(\mathbf{x}) U^\top W_\tau \mathbf{Y} - \rho^{-1}\mathbf{r}_k^\top(\mathbf{x}) U^\top W_\tau U (\rho\Sigma_k^{-1} + U^\top W_\tau U)^{-1} U^\top W_\tau \mathbf{Y}.$$

According to the definition in Proposition 1, we notice that

$$U^\top W_\tau U = \Lambda_\tau,$$

and

$$\Lambda_\tau^{-1} U^\top W_\tau \mathbf{Y} = \bar{\mathbf{Y}}_\tau.$$

Therefore, we obtain

$$\hat{f}_\tau(\mathbf{x}) = \rho^{-1}\left[\mathbf{r}_k^\top(\mathbf{x})\Lambda_\tau\bar{\mathbf{Y}}_\tau - \mathbf{r}_k^\top(\mathbf{x})\Lambda_\tau(\rho\Sigma_k^{-1}+\Lambda_\tau)^{-1}\Lambda_\tau\bar{\mathbf{Y}}_\tau\right].$$

By using the Woodbury identity to compute $(\rho\Sigma_k^{-1}+\Lambda_\tau)^{-1}$, we have

$$(\rho\Sigma_k^{-1}+\Lambda_\tau)^{-1} = \Lambda_\tau^{-1} - \Lambda_\tau^{-1}(\rho^{-1}\Sigma_k+\Lambda_\tau^{-1})^{-1}\Lambda_\tau^{-1},$$

which further gives the conclusion in the proposition.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

Binois, M., R. B. Gramacy, and M. Ludkovski. 2016. "Practical Heteroskedastic Gaussian Process Modeling for Large Simulation Experiments". *arXiv preprint arXiv:1611.05902*.

Cressie, N. 2015. *Statistics for spatial data*. John Wiley & Sons.

Efron, B. 1991. "Regression Percentiles Using Asymmetric Squared Error Loss". *Statistica Sinica* 1 (1): 93–125.

Henderson, S. G., and B. L. Nelson. 2006. *Handbooks in Operations Research and Management Science: Simulation*, Volume 13. NORTH-HOLLAND.

Koenker, R. 2005. *Quantile Regression*. Number 38. Cambridge university press.

Kuan, C.-M., J.-H. Yeh, and Y.-C. Hsu. 2009. "Assessing Value at Risk with Care, The Conditional Autoregressive Expectile Models". *Journal of Econometrics* 150 (2): 261–270.

Newey, W. K., and J. L. Powell. 1987. "Asymmetric Least Squares Estimation and Testing". *Econometrica: Journal of the Econometric Society*:819–847.

Plumlee, M., and R. Tuo. 2014. "Building Accurate Emulators for Stochastic Simulations via Quantile Kriging". *Technometrics* 56 (4): 466–473.

Schölkopf, B., R. Herbrich, and A. J. Smola. 2001. "A Generalized Representer Theorem". In *Computational learning theory*, 416–426. Springer.

Sun, L., L. J. Hong, and Z. Hu. 2014. "Balancing Exploitation And Exploration In Discrete Optimization Via Simulation Through A Gaussian Process-based Search". *Operations Research* 62 (6): 1416–1438.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl. 2013. "Using Proper Divergence Functions to Evaluate Climate Models". *SIAM/ASA Journal on Uncertainty Quantification* 1 (1): 522–534.

Wahba, G. 1990. *Spline Models for Observational Data*, Volume 59. Siam.

Xie, W., B. L. Nelson, and R. R. Barton. 2014. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* 62 (6): 1439–1452.

AUTHOR BIOGRAPHIES

QIONG ZHANG is an Assistant Professor of statistics at Virginia Commonwealth University, Richmond, VA. She holds Ph.D. degree in statistics from University of Wisconsin-Madison. Her research interests include computer experiments, uncertainty quantification and spatial and spatial-temporal modeling. She is a member of ASA and INFORMS. Her email address is qzhang4@vcu.edu.

WEI XIE is an assistant professor in the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. She received her M.S. and Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests are in computer simulation, risk management and data analytics. Her email address is xiew3@rpi.edu and her web page is http://homepages.rpi.edu/~xiew3/.