Enhancing Stochastic Kriging for Queueing Simulation with Stylized Models

Haihui Shen[1], L. Jeff Hong[1], and Xiaowei Zhang[2]

[1]Department of Management Sciences, City University of Hong Kong, Kowloon Tong, Hong Kong
[2]Department of Industrial Engineering and Decision Analytics, HKUST, Clear Water Bay, Hong Kong

Abstract

Stochastic kriging is a popular metamodeling technique to approximate computationally expensive simulation models. However, it typically treats the simulation model as a black box in practice and often fails to capture the highly nonlinear response surfaces that arise from queueing simulations. We propose a simple, effective approach to improve the performance of stochastic kriging by incorporating stylized queueing models which contain useful information about the shape of the response surface. We provide several statistical tools to measure usefulness of the incorporated stylized models. We show that even a relatively crude stylized model can improve the prediction accuracy of stochastic kriging substantially.

Key words: stochastic kriging; metamodel; queueing simulation; stylized queueing model

1 Introduction

Queueing models are widely used to facilitate decision making in a great variety of areas, including manufacturing, logistics, supply chain management, telecommunication, health care, finance, etc. However, they generally do not admit analytical expressions except those that are highly stylized such as Erlang's loss systems and Jackson networks (Asmussen 2008). Instead, simulation is extensively adopted to analyze and predict the behavior of complex queueing models that arise from large-scale stochastic systems in real world applications. The popularity of queueing simulation stems from its modeling flexibility, allowing the users to incorporate arbitrarily fine details of the system into the model and estimate virtually any performance measure of interest. Nevertheless, typical queueing simulations are computationally expensive to execute, especially if the performance measures of interest are steady-state quantities or if the systems are heavily utilized (Whitt 1989, Asmussen 1992). The computational inefficiency severely restricts usefulness of queueing simulation in settings such as real-time decision making and system optimization.

In order to alleviate this inadequacy, *metamodeling* has been actively developed in the simulation community; see Barton (1998) for a review. A metamodel, or a model of the simulation model,

aims to characterize the performance measure of the simulation model, i.e. response surface, as a function of the design variables. It is often built via proper interpolation of the simulation outputs at a small number of carefully chosen design points. A metamodel runs much faster than the original simulation model in general, and it yields deterministic outputs rather than stochastic. Hence, it can be used in lieu of the true response surface to efficiently search for the optimal values of the design variables, even in real time.

Kriging-type metamodels have recently become popular in the simulation literature, thanks to their tractability, ease of use, and capability of providing good global fit over the value range of the design variables and capturing moderate heteroskedasticity of the response surfaces. Kriging originated in geostatistics (Matheron 1963) and was later successfully adopted in the design and analysis of computer experiments (DACE) community to fit deterministic simulation models (Sacks et al. 1989); see Kleijnen (2009) for a review. In this paper, we focus on stochastic kriging which was proposed by Ankenman et al. (2010) to account for the intrinsic uncertainty in stochastic simulation that results from the random simulation noise. This metamodel has been used successfully for uncertainty quantification in stochastic simulation (Xie et al. 2014) and simulation optimization (Scott et al. 2011, Sun et al. 2014).

Simulation metamodels including stochastic kriging treat the simulation model to approximate as a black box in general, discarding its internal details and structural properties of the response surface. This issue may become severe when the response surface is highly nonlinear or even exhibits "exploding" behavior, which is often the case for queueing simulation if the simulated queue is near saturation with a high utilization. Our solution to this issue takes advantage of stylized queueing models which are highly analytically tractable due to their greatly simplifying assumptions. Albeit not good at quantitative prediction, stylized models can capture the essential dynamics of the queueing system, facilitating the development of managerial insights. For example, they may help the users to identify the bottleneck of a queueing network and its saturation mechanism.

The central idea of this paper is to *use the stylized queueing model to capture the highly nonlinear trend of the response surface, use regression that is linear in the unknown coefficients to adjust both the scaling factor of the stylized queueing model and the mean level of the "detrended" surface, and use the spatial correlation inherent in stochastic kriging to correct the remaining bias.* Notice that stylized queueing models generally have analytical solutions or simple numerical solutions for the performance measure of interest, so their computational complexity is negligible relative to the simulation model. By incorporating stylized queueing models, we effectively extract the valuable structural information about the response surface from them and transform the queueing simulation model from a black box to a gray box. We will demonstrate that the gray-box perspective greatly enhances the performance of stochastic kriging in the context of queueing simulation and significantly accelerate the process of solving the associated simulation optimization problems. Other contributions of this paper include developing several statistical tools to measure usefulness and effectiveness of a stylized queueing model in the proposed metamodel.

Incorporating context-specific information to improve the prediction accuracy of metamodels for

queueing simulation is not new; see Cheng and Kleijnen (1999) and Yang et al. (2007). Specifically, they assume that the "trend" of the response surface consists of two factors, one of which accounts for the exploding behavior of saturated queues. However, this factor requires that the users know exactly the saturation point where the explosion occurs, which is not necessarily the case for complex queueing networks and even prohibitively difficult if the design variable of interest is multidimensional. Moreover, their metamodels are based on low-order polynomial regressions, which tend to provide good fit only locally, rather than the kind of robust global fit that stochastic kriging aims for. In the same vein, Lin et al. (2016) recently proposed to leverage an analytical model to enhance kernel regression, another popular metamodeling technique. In their work, the simulation outputs are adjusted by the outputs of the analytical model before being used in kernel regression. The net effect is that the predicted responses at locations that are distant from (resp., close to) the design points are basically determined by the analytical (resp., simulation) model. By contrast, the use of analytical models in our paper is to detrend the response surface so that the residuals can be better fitted by a stationary random field. A notable result of our different treatment is that both the analytical and simulation models have a nonnegligible impact in the prediction of the responses at all locations.

There are other ways to enhance stochastic kriging as well by incorporating auxiliary information. One approach is to leverage the gradient information of the response surface, provided that it can be acquired along with the observations of the surface itself easily; see Morris et al. (1993) and Mitchell et al. (1994) in the DACE setting, and Chen et al. (2013) and Qu and Fu (2014) in the stochastic simulation settings. Another main approach is to assume that a coarser but faster simulation model of the same system is available in addition to the original expensive simulation model, and then leverage the simulation outputs from both models and the correlations between them to refine the prediction; see Kennedy and O'Hagan (2000) and Forrester et al. (2007). However, both of the approaches above adopt a black-box perspective and generally do not take into account the structural information that is necessary to represent the highly heteroskedastic response surfaces of queueing simulation. Moreover, our approach is orthogonal to theirs in a sense and can be used in combination with them to achieve further enhancement.

Our approach to enhancing stochastic kriging can be viewed as a means for improving the trend modeling. Other aspects that one can investigate to improve the prediction accuracy of stochastic kriging include experiment design (e.g., design point placement and simulation budget allocation) and choice of the covariance function. Design of computer experiments is an research area of great importance in its own right and we refer to Santner et al. (2003, §5 and §6) for a general exposition on the subject. In particular, the experiment design proposed by Ankenman et al. (2010) which assumes a constant trend term in stochastic kriging and attempts to minimize the integrated mean squared error of prediction over the entire design space can be potentially extended to our setting without essential difficulty.

The choice of the covariance function, on the other hand, can be viewed formally as a model selection problem, and thus statistical tools such as information criteria or cross validation can

be readily applied; see Rasmussen and Williams (2006, §5). One can also choose the covariance function based on domain knowledge such as the smoothness of the response surface, since the level of differentiability is uniquely determined by that of the covariance function; see Xie et al. (2010). We remark that the three aspects – trend modeling, experiment design, and choice of covariance function – are essentially orthogonal but equally important for enhancing stochastic kriging. The approach that the present paper follows represents a simple way to improve trend modeling, and any improvement made to the other two aspects can and should be incorporated in practice to further boost the prediction accuracy of stochastic kriging.

The remainder of this paper is organized as follows. In §2 we build the framework for stylized-model enhanced stochastic kriging. In §3 we present several statistical measures for evaluating stylized queueing models. In §4 we demonstrate through an illustrative example the benefits of stylized models and validity of measures. We also compare our stylized-model enhanced stochastic kriging with another approach that enhances stochastic kriging by leveraging gradient information. In §5 we propose a few simple approaches for constructing stylized models for a general class of queueing networks. We study two real-world applications in §6 and conclude in §7. The Appendix collects some of the technique results.

2 The Framework

In this section, we give an overview of the stochastic kriging metamodel proposed in Ankenman et al. (2010), discuss its deficiency in practice, and introduce our approach for incorporating stylized queueing models in stochastic kriging.

2.1 Stochastic Kriging

Metamodeling is concerned with fitting an unknown *deterministic* response surface $\mathcal{Y}(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal$ denotes the design variables of the simulation model. For example, $\boldsymbol{x}$ may represent the number of servers and their service capacities, while $\mathcal{Y}(\boldsymbol{x})$ the steady-state mean waiting time of the system. The kriging method expresses $\mathcal{Y}(\boldsymbol{x})$ as

$$\mathcal{Y}(\boldsymbol{x}) \coloneqq \boldsymbol{f}(\boldsymbol{x})^\intercal \boldsymbol{\beta} + \mathcal{M}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is a vector of known functions of $\boldsymbol{x}$, $\boldsymbol{\beta}$ is a vector of unknown parameters of compatible dimension, and $\mathcal{M}$ is a *realization* of a mean zero random field, which is randomly sampled from a space of functions mapping $\mathbb{R}^d \mapsto \mathbb{R}$. A typical example of $\boldsymbol{f}(\boldsymbol{x})$ is basis functions, such as polynomials. The metamodel (1) is called "universal kriging"; in particular, if the "trend" $\boldsymbol{f}(\boldsymbol{x})^\intercal \boldsymbol{\beta} \equiv \beta_0$ is a constant, it is called "ordinary kriging". In the context of stochastic simulation, $\mathcal{Y}(\boldsymbol{x})$ is observed with random noise. Therefore, the stochastic kriging metamodel assumes that the simulation output on the j-th replication at design point $\boldsymbol{x}$ is

$$Y_j(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\intercal \boldsymbol{\beta} + \mathcal{M}(\boldsymbol{x}) + \varepsilon_j(\boldsymbol{x}), \tag{2}$$

where $\varepsilon_1(\boldsymbol{x}), \varepsilon_2(\boldsymbol{x}), \ldots$ are the simulation errors. Suppose that the simulation model is executed at design points $\boldsymbol{x}_i$ with n_i simulation replications, $i = 1, \ldots, k$. Define

$$\bar{Y}(\boldsymbol{x}_i) \coloneqq \frac{1}{n_i} \sum_{j=1}^{n_i} Y_j(\boldsymbol{x}_i) \quad \text{and} \quad \bar{\varepsilon}(\boldsymbol{x}_i) \coloneqq \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_j(\boldsymbol{x}_i).$$

Then, the metamodel (2) can be rewritten as

$$\bar{\boldsymbol{Y}} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{M} + \bar{\boldsymbol{\varepsilon}},$$

where $\bar{\boldsymbol{Y}} \coloneqq (\bar{Y}(\boldsymbol{x}_1), \ldots, \bar{Y}(\boldsymbol{x}_k))^\intercal$, $\bar{\boldsymbol{\varepsilon}} \coloneqq (\bar{\varepsilon}(\boldsymbol{x}_1), \ldots, \bar{\varepsilon}(\boldsymbol{x}_k))^\intercal$, $\boldsymbol{F} \coloneqq (\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_k))^\intercal$, and $\boldsymbol{M} \coloneqq (\mathcal{M}(\boldsymbol{x}_1), \ldots, \mathcal{M}(\boldsymbol{x}_k))^\intercal$. Further, let $\boldsymbol{\Gamma}$ be the $k \times k$ covariance matrix of $\boldsymbol{M}$, i.e. $\boldsymbol{\Gamma}_{ij} = \mathrm{Cov}[\mathcal{M}(\boldsymbol{x}_i), \mathcal{M}(\boldsymbol{x}_j)]$ for $i, j = 1, \ldots, k$. Let $\boldsymbol{\gamma}(\boldsymbol{x}_0)$ be the $k \times 1$ vector $(\mathrm{Cov}[\mathcal{M}(\boldsymbol{x}_0), \mathcal{M}(\boldsymbol{x}_1)], \ldots, \mathrm{Cov}[\mathcal{M}(\boldsymbol{x}_0), \mathcal{M}(\boldsymbol{x}_k)])^\intercal$. Let $\boldsymbol{\Sigma} = \mathrm{Cov}[\bar{\boldsymbol{\varepsilon}}, \bar{\boldsymbol{\varepsilon}}]$ be the $k \times k$ covariance matrix of $\bar{\boldsymbol{\varepsilon}}$, i.e. $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}[\bar{\varepsilon}(\boldsymbol{x}_i), \bar{\varepsilon}(\boldsymbol{x}_j)]$ for $i, j = 1, \ldots, k$. The following assumption is usually imposed for stochastic kriging.

Assumption 1. *The random field $\mathcal{M}$ is a second-order stationary Gaussian process with mean 0. More specifically, $\mathbb{E}[\mathcal{M}(\boldsymbol{x})] \equiv 0$ and $\mathrm{Cov}[\mathcal{M}(\boldsymbol{x}), \mathcal{M}(\boldsymbol{x}')] = \tau^2 R(\boldsymbol{x} - \boldsymbol{x}'; \boldsymbol{\theta})$, where τ^2 is the variance of $\mathcal{M}(\boldsymbol{x})$ for all $\boldsymbol{x}$ and R is the correlation function that depends only on $\boldsymbol{x} - \boldsymbol{x}'$ and some unknown parameters $\boldsymbol{\theta}$. Moreover, R satisfies $R(\boldsymbol{0}; \boldsymbol{\theta}) = 1$ and $R(\boldsymbol{x} - \boldsymbol{x}'; \boldsymbol{\theta}) \to 0$ as $\|\boldsymbol{x} - \boldsymbol{x}'\| \to \infty$. The simulation errors $\varepsilon_1(\boldsymbol{x}), \varepsilon_2(\boldsymbol{x}), \ldots$ are independent and identically distributed with normal distribution $\mathcal{N}(0, \sigma^2(\boldsymbol{x}))$, and independent of $\mathcal{M}$.*

Remark 1. In all the numerical examples of this paper, we assume a Gaussian correlation function of form $R(\boldsymbol{x} - \boldsymbol{x}'; \boldsymbol{\theta}) = \exp(-\sum_{i=1}^{d} \theta_i |x_i - x_i'|^2)$ with $\theta_i > 0$ for $i = 1, \ldots, d$; see Stein (1999, §2.7) for more types of correlation functions.

We are interested in predicting the response $\mathcal{Y}(\boldsymbol{x}_0)$ at an arbitrary point $\boldsymbol{x}_0$. It can be shown with a similar derivation in Stein (1999, §1.2) that under Assumption 1, the best unbiased predictor that minimizes the mean squared error (MSE) of prediction is

$$\widehat{\mathcal{Y}}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0)^\intercal \boldsymbol{\beta} + \boldsymbol{\gamma}(\boldsymbol{x}_0)^\intercal \left(\boldsymbol{\Gamma} + \boldsymbol{\Sigma}\right)^{-1} \left(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta}\right),$$

and the optimal MSE is

$$\mathrm{MSE}^*(\boldsymbol{x}_0) = \tau^2 - \boldsymbol{\gamma}(\boldsymbol{x}_0)^\intercal \left(\boldsymbol{\Gamma} + \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\gamma}(\boldsymbol{x}_0), \tag{3}$$

provided that $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}(\boldsymbol{x}_0)$, and $\boldsymbol{\Sigma}$ are known.

2.2 Stylized-Model Enhanced Stochastic Kriging for Queueing Simulation

Despite its general form, in applications $\boldsymbol{f}(\boldsymbol{x})$ is mostly taken as a constant, i.e. $\boldsymbol{f}(\boldsymbol{x})^\intercal \boldsymbol{\beta} \equiv \beta_0$. For instance, such specification is recommended in Ankenman et al. (2010); see also Sacks et al. (1989) and Kennedy and O'Hagan (2001). A main reason, in addition to the obvious simplicity, is

that $\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{\beta}$ represents the user's knowledge about the shape of the response surface. For complex simulation models, it is often difficult to acquire such meaningful information in advance. To avoid introducing spurious constituent functions to the metamodel, it is generally preferable to take a black-box viewpoint and use a constant trend, unless there is actual prior information that suggests otherwise. Following the naming convention in the kriging literature, we call stochastic kriging with a constant trend *ordinary stochastic kriging* (OSK).

However, the response surfaces in the queueing simulation setting are highly nonlinear and highly heteroskedastic in general, and often exhibit exploding behavior as the utilization of the queue increases. Given the high computational cost, both the number of design points at which the simulation model is executed and the number of simulation replications are commonly limited. In this case, the simulation outputs are insufficient to reveal the true shape of the surface, leading to inaccurate predictions. To illustrate, consider the steady-state expected number of customers (including the one in service) $\mathcal{Y}(x)$ of an $M/M/1$ queue with utilization $x \in (0,1)$. It is well known that $\mathcal{Y}(x) = x/(1-x)$, and clearly the surface explodes as x approaches 1. In order to highlight the deficiency of using a constant trend, we assume that the simulation is "noiseless" and $\mathcal{Y}(x)$ can be computed without error for any arbitrary x. OSK is then reduced to traditional ordinary kriging in this case. We compute $\mathcal{Y}(x)$ at 5 design points $x = 0.1, 0.3, 0.5, 0.7, 0.9$ and take a Gaussian correlation function of the form $R(x;\theta) = e^{-\theta x^2}$ for the ordinary kriging metamodel. Figure 1 shows that the predicted surface of ordinary kriging is reasonably good for most of the value range of x. However, it fails to capture the unboundedness of $\mathcal{Y}(x)$ when the queue is saturated, which is of greater interest for decision makers. (Notice that the predicted surface for $x \in [0.9, 1)$ appears to be stabilized whereas the true surface begins to explode.)

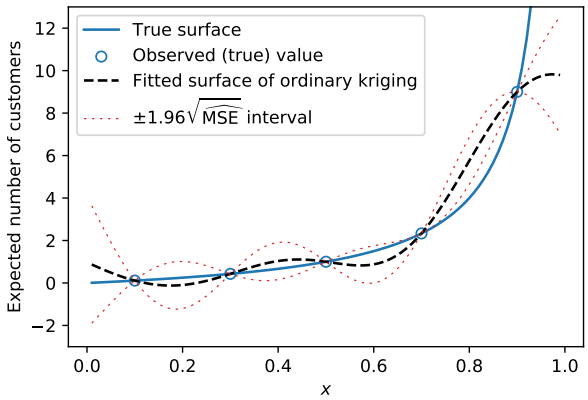

Figure 1: Ordinary Kriging for $M/M/1$ Queue.

In addition to failing to capture the non-constant trend, OSK tends to overestimate the marginal variance of the constituent Gaussian process in order to compensate the possibly large variation in the response surface, which should have been characterized by the trend term. The incorrect estimation of the model variability may become a significant issue both in statistical inference and in simulation optimization algorithms which often use the variance information to determine the

exploitation-exploration trade-off (Sun et al. 2014). If the trend can be reasonably captured, the detrended surface would have substantially less variation, and thus is more suitable to be modeled as a second-order stationary Gaussian process. See also §3.2 for more discussion on the issue.

To alleviate the inadequacy of OSK, we incorporate "informative" functions into $\boldsymbol{f}(\boldsymbol{x})$. We argue that basis functions such as polynomials, splines, radial basis function, etc. are not particularly suitable for the queueing simulation setting, because they lack domain knowledge of the problem context and do not capture the exploding behavior. And it can be very difficult to specify proper basis functions when the design variable $\boldsymbol{x}$ is multidimensional. In addition, as the number of basis functions increases, the users need to address the issue of overfitting and the numerical challenge in the parameter estimation caused by high-dimensional numerical optimization problems.

Instead, we take advantage of stylized queueing models which can represent the essential structure of the complex queueing network being simulated, and meanwhile admit analytical solutions or simple numerical solutions that can be used as informative functions in the metamodel. Indeed, we shall discuss in §5 and demonstrate in §6.1 and §6.2 that these stylized queueing models can be easily constructed for a large class of queueing networks. For the time being, suppose that we have built a proper stylized queueing model with response $q(\boldsymbol{x})$. Then, the metamodel we propose for queueing simulation has the same form as (2) with $\boldsymbol{f}(\boldsymbol{x}) = (1, q(\boldsymbol{x}))^\intercal$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\intercal$. It is straightforward to extend the formulation to the case of multiple stylized queueing models, but we focus on the simple case. We call this metamodel *stylized-model enhanced stochastic kriging* (SESK). Clearly, the purpose of $q(\boldsymbol{x})$ is to capture the trend of the true response surface, whereas the coefficients β_1 and β_0 are used to represent the scaling factor of $q(\boldsymbol{x})$ and the mean level of the detrended surface via linear regression, respectively.

2.3 Maximum Likelihood Estimation

To apply SESK in practice, the parameters $\boldsymbol{\beta}$, τ^2, $\boldsymbol{\theta}$, and $\boldsymbol{\Sigma}$ need to be estimated. In this section, we extend the maximum likelihood estimation (MLE) developed in Ankenman et al. (2010) for OSK.

First, $\boldsymbol{\Sigma}$ represents the intrinsic uncertainty of the simulation, and its estimation can be separated from the other parameters. Due to the heteroskedasticity in queueing simulation, it is difficult to construct a parametric model for $\boldsymbol{\Sigma}$. Instead, we estimate $\boldsymbol{\Sigma}$ by the sample variances of the simulation outputs, i.e. $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\widehat{\sigma}^2(\boldsymbol{x}_1)/n_1, \ldots, \widehat{\sigma}^2(\boldsymbol{x}_k)/n_k)$, where

$$\widehat{\sigma}^2(\boldsymbol{x}_i) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_j(\boldsymbol{x}_i) - \bar{Y}(\boldsymbol{x}_i))^2. \tag{4}$$

Theorem 1 of Ankenman et al. (2010) shows that using this estimator of $\boldsymbol{\Sigma}$ does not introduce prediction bias.

Clearly, the simulation outputs $\bar{\boldsymbol{Y}}$ are multivariate normal under Assumption 1. Hence, assum-

ing $\boldsymbol{\Sigma}$ is known, the log-likelihood function of $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{\Gamma}(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}\right| - \frac{1}{2}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta})^\intercal \left[\boldsymbol{\Gamma}(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}\right]^{-1} (\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta}), \qquad (5)$$

where $|\cdot|$ denotes the determinant of a matrix and we write $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\tau^2, \boldsymbol{\theta})$ to stress the dependence on the parameters.

The first-order conditions for maximizing $\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ can be easily obtained by applying standard results for matrix calculus and we omit the details; see Stein (1999, §6.4) for discussion on related numerical methods.

In summary, a stochastic kriging metamodel is constructed as follows:

(i) Estimate $\boldsymbol{\Sigma}$ as $\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}(\widehat{\sigma}^2(\boldsymbol{x}_1)/n_1, \ldots, \widehat{\sigma}^2(\boldsymbol{x}_k)/n_k)$, where $\widehat{\sigma}^2(\boldsymbol{x}_i)$ is given by (4).

(ii) Using $\widehat{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$, maximize $\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ in (5) to obtain the estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$. Set $\widehat{\boldsymbol{\Gamma}} \coloneqq \boldsymbol{\Gamma}(\widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$, and $\widehat{\boldsymbol{\gamma}}(\boldsymbol{x}_0) \coloneqq \boldsymbol{\gamma}(\widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$.

(iii) Predict $\mathcal{Y}(\boldsymbol{x}_0)$ via the plug-in predictor,

$$\widehat{\widehat{\mathcal{Y}}}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0)^\intercal \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\gamma}}(\boldsymbol{x}_0)^\intercal \left(\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1} \left(\bar{\boldsymbol{Y}} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}\right),$$

with MSE estimator

$$\widehat{\mathrm{MSE}}(\boldsymbol{x}_0) = \widehat{\tau}^2 - \widehat{\boldsymbol{\gamma}}(\boldsymbol{x}_0)^\intercal \left(\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1} \widehat{\boldsymbol{\gamma}}(\boldsymbol{x}_0) + \boldsymbol{\delta}^\intercal \left[\boldsymbol{F}^\intercal \left(\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1} \boldsymbol{F}\right]^{-1} \boldsymbol{\delta},$$

where $\boldsymbol{\delta} = \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{F}^\intercal \left(\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1} \widehat{\boldsymbol{\gamma}}(\boldsymbol{x}_0)$; see Stein (1999, §1.5) for a similar derivation.

3 Measures for the Stylized Model

Obviously, multiple stylized queueing models can be constructed for SESK for a given queueing simulation experiment. Two natural but important questions then arise. (i) How do we test whether a stylized model indeed provides useful information on the response surface? (ii) How do we select a stylized model from a set of candidate models? The two questions are about the usefulness and effectiveness of a stylized model, respectively. Since the true surface is unknown, these questions cannot be addressed by comparing it with the predicted surface, but based on the statistical evidence provided by the simulation outputs at the design points. In this section, we devise a hypothesis test to address the usefulness and propose a new statistic to measure the effectiveness.

3.1 Z-Test for Usefulness

We propose the following hypothesis test for usefulness of a given stylized model $q(x)$ in SESK:

- null hypothesis H_0: $\beta_1 = 0$;
- alternative hypothesis H_1: $\beta_1 \neq 0$.

If H_0 is rejected, then $q(x)$ is useful for capturing the trend of the response surface.

Throughout this section, we assume that $\boldsymbol{\Sigma}$ is given. For notational simplicity, let $\boldsymbol{\psi} \coloneqq (\tau^2, \boldsymbol{\theta}^\intercal)^\intercal$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\psi}) \coloneqq \boldsymbol{\Gamma}(\tau^2, \boldsymbol{\theta}) + \boldsymbol{\Sigma}$. Let m be the size of $\boldsymbol{\psi}$, i.e. $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^\intercal$. Moreover, we use $\boldsymbol{\beta}^0$ and $\boldsymbol{\psi}^0$ to denote the (unknown) true value of $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, respectively; let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\psi}}$ denote their respective maximum likelihood (ML) estimators given $\boldsymbol{\Sigma}$.

In order to construct a test statistic, we need to derive the large-sample asymptotic distribution of the ML estimator $\widehat{\beta}_1$. There are two large-sample asymptotic regimes for MLE in spatial statistics, i.e. increasing-domain regime and fixed-domain regime (Zhang and Zimmerman 2005). The former assumes that the minimum distance between the design points is bounded away from zero and the sampling domain is unbounded. By contrast, the latter assumes that the design points are taken more and more densely from a fixed and bounded domain. We adopt the increasing-domain regime (Assumption 2), because the ML estimators in the fixed-domain regime may be inconsistent even for some widely used correlation functions (Zhang 2004). We impose in Assumption 3 certain regularity conditions on $\boldsymbol{\Omega}$ and $\boldsymbol{F}$ in order to obtain the consistency and asymptotic normality of the ML estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}})$ in the increasing-domain regime. Both Assumptions 2 and 3 are standard (Mardia and Marshall 1984) and we briefly discuss their applicability in Remarks 2 and 3, respectively.

Assumption 2. *The design points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ form a regular lattice and there exists a constant $c > 0$ such that $\|\boldsymbol{x}_r - \boldsymbol{x}_s\| \geq c$ for all $r, s = 1, \ldots, k$, as $k \to \infty$. For $\boldsymbol{h} \in \mathbb{Z}^d$, letting $\sigma_{\boldsymbol{h}} \coloneqq \mathrm{Cov}[\mathcal{M}(\boldsymbol{x} + \boldsymbol{h}), \mathcal{M}(\boldsymbol{x})]$ is absolutely summable over $\mathbb{Z}^d$, i.e. $\sum_{\boldsymbol{h} \in \mathbb{Z}^d} |\sigma_{\boldsymbol{h}}| < \infty$. Moreover, both $\sigma_{\boldsymbol{h},i}$ and $\sigma_{\boldsymbol{h},ij}$ are absolutely summable over $\mathbb{Z}^d$ for all $i, j = 1, \ldots, m$, where $\sigma_{\boldsymbol{h},i} = \partial \sigma_{\boldsymbol{h}} / \partial \psi_i$, and $\sigma_{\boldsymbol{h},ij} = \partial^2 \sigma_{\boldsymbol{h}} / \partial \psi_i \partial \psi_j$.*

Remark 2. It can be verified easily that the power exponential correlation function of the form $R(\boldsymbol{x} - \boldsymbol{x}'; \boldsymbol{\theta}) = \exp(-\sum_{i=1}^d \theta_i |x_i - x_i'|^\alpha)$ for $\alpha \in (0, 2]$ satisfies the absolute summability conditions in Assumption 2.

Assumption 3. *Suppose the following regularity conditions on $\boldsymbol{\Omega}$ and $\boldsymbol{F}$:*

(i) $\boldsymbol{\Omega}$ is positive definite and twice differentiable with respect to $\boldsymbol{\psi}$ for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a compact set containing the true value $\boldsymbol{\psi}^0$ as an interior point;

(ii) $\boldsymbol{F}$ has full rank and $(\boldsymbol{F}^\intercal \boldsymbol{F})^{-1}$ converges to a zero matrix as $k \to \infty$;

(iii) for all $i, j = 1, \ldots, m$, $t_{ij} / (t_{ii} t_{jj})^{1/2}$ converges as $k \to \infty$ to a finite limit a_{ij}, where $t_{ij} \coloneqq \mathrm{tr}(\boldsymbol{\Omega}^{-1} (\partial \boldsymbol{\Omega}^{-1} / \partial \psi_i) \boldsymbol{\Omega}^{-1} (\partial \boldsymbol{\Omega}^{-1} / \partial \psi_j))$, and $A \coloneqq (a_{ij})_{i,j=1,\ldots,m}$ is a nonsingular matrix;

(iv) $k(\boldsymbol{F}^\intercal \boldsymbol{\Omega}(\boldsymbol{\psi})^{-1} \boldsymbol{F})^{-1}$ converges to a finite limit uniformly for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ as $k \to \infty$.

Remark 3. Condition (i) of Assumption 3 is satisfied for many popular correlation functions including the power exponential correlation function. Notice that $\boldsymbol{F} = (\mathbf{1}, \boldsymbol{q}) \in \mathbb{R}^{k\times 2}$, where $\mathbf{1} \in \mathbb{R}^k$ is the column vector of ones and $\boldsymbol{q} = (q(\boldsymbol{x}_1), \ldots, q(\boldsymbol{x}_k))^\intercal$. Hence, the first part of condition (ii) is trivially satisfied if the stylized model $q(\boldsymbol{x})$ is not a constant. The second part of condition (ii) holds if the smallest eigenvalue of $\boldsymbol{F}^\intercal\boldsymbol{F}$ tends to infinity as $k \to \infty$. Simple linear algebra reveals that this is equivalent to $\|\boldsymbol{q}\|^2 + k - \sqrt{(\|\boldsymbol{q}\|^2 - k)^2 + 4\boldsymbol{q}^\intercal\mathbf{1}} \to \infty$ as $k \to \infty$. Conditions (iii) and (iv) ensure that the Fisher information matrix of the ML estimator is well behaved in the limit.

Now we derive some large-sample asymptotic properties of the ML estimators. Define for any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, $\boldsymbol{V}^{(k)}(\boldsymbol{\psi}) \coloneqq k(\boldsymbol{F}^\intercal\boldsymbol{\Omega}(\boldsymbol{\psi})^{-1}\boldsymbol{F})^{-1}$ and $\boldsymbol{V}(\boldsymbol{\psi}) \coloneqq \lim_{k\to\infty} \boldsymbol{V}^{(k)}(\boldsymbol{\psi})$. Notice that $\boldsymbol{V}^{(k)}(\boldsymbol{\psi}) \in \mathbb{R}^{2\times 2}$ for any k, and $\boldsymbol{V}(\boldsymbol{\psi}) \in \mathbb{R}^{2\times 2}$. What's more, $\boldsymbol{V}(\boldsymbol{\psi})$ is positive definite and continuous in $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, because the convergence of $\boldsymbol{V}^{(k)}(\boldsymbol{\psi})$ is uniform for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$. The main result of the asymptotic distribution of the ML estimator $\widehat{\boldsymbol{\beta}}$ is stated in the following Theorem 1.

Theorem 1. *Let $\boldsymbol{B} \in \mathbb{R}^{n\times 2}$ be a matrix of rank n ($n \le 2$). Then, under Assumptions 1 - 3,*

$$\sqrt{k}[\boldsymbol{B}\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}})\boldsymbol{B}^\intercal]^{-\frac{1}{2}}(\boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{B}\boldsymbol{\beta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{6}$$

as $k \to \infty$, where $[\boldsymbol{B}\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}})\boldsymbol{B}^\intercal]^{\frac{1}{2}}$ denotes the square root of $\boldsymbol{B}\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}})\boldsymbol{B}^\intercal$, $\mathbf{0}$ is the $n \times 1$ zero vector, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and $\xrightarrow{d}$ means "converges in distribution".

The proof of Theorem 1 is based on the following Lemma 1, which is a direct application of the Theorems 1 - 3 in Mardia and Marshall (1984).

Lemma 1. *Under Assumptions 1 - 3,*

$$\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0 \quad \textit{and} \quad \sqrt{k}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V}(\boldsymbol{\psi}^0)), \tag{7}$$

as $k \to \infty$, where $\xrightarrow{p}$ means "converges in probability".

Proof of Theorem 1. We first prove that as $k \to \infty$,

$$\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}}) \xrightarrow{p} \boldsymbol{V}(\boldsymbol{\psi}^0). \tag{8}$$

This is equivalent to show that for any $\epsilon, \delta > 0$, there exist $M < \infty$ for which

$$\Pr(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \ge \epsilon) < \delta, \quad k > M,\ i, j = 1, 2. \tag{9}$$

From the first part of (7) of Lemma 1, $\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$ as $k \to \infty$. Since $\boldsymbol{V}(\boldsymbol{\psi})$ is continuous in $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, the continuous mapping theorem (Billingsley 1999, §1.2) indicates that conditionally on $\widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}$, $\boldsymbol{V}(\widehat{\boldsymbol{\psi}}) \xrightarrow{p} \boldsymbol{V}(\boldsymbol{\psi}^0)$ as $k \to \infty$. Hence, there exists $M_1 < \infty$ for which

$$\Pr\left(|V_{ij}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \ge \frac{\epsilon}{2} \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) < \frac{\delta}{2}, \quad k > M_1,\ i, j = 1, 2. \tag{10}$$

Since $\boldsymbol{V}^{(k)}(\boldsymbol{\psi}) \to \boldsymbol{V}(\boldsymbol{\psi})$ uniformly for $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, there exists $M_2 < \infty$ for which

$$|V_{ij}^{(k)}(\boldsymbol{\psi}) - V_{ij}(\boldsymbol{\psi})| < \frac{\epsilon}{2}, \quad k > M_2,\ \boldsymbol{\psi} \in \boldsymbol{\Psi},\ i,j = 1,2. \tag{11}$$

Since $\widehat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$ and $\boldsymbol{\psi}^0$ is an interior point of $\boldsymbol{\Psi}$, there exist $M_3 < \infty$ for which

$$\Pr(\widehat{\boldsymbol{\psi}} \notin \boldsymbol{\Psi}) < \frac{\delta}{2}, \quad k > M_3. \tag{12}$$

Consequently, for any $i,j = 1,2$ and $k > M \coloneqq \max\{M_1, M_2, M_3\}$,

$$\begin{aligned}
\Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) &= \Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\widehat{\boldsymbol{\psi}}) + V_{ij}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) \\
&\leq \Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\widehat{\boldsymbol{\psi}})| + |V_{ij}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) \\
&\leq \Pr\left(\frac{\epsilon}{2} + |V_{ij}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) \\
&< \frac{\delta}{2},
\end{aligned} \tag{13}$$

where the second inequality follows from (11) and the third from (10). Therefore,

$$\begin{aligned}
&\Pr(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon) \\
=&\Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \notin \boldsymbol{\Psi}\right) \Pr(\widehat{\boldsymbol{\psi}} \notin \boldsymbol{\Psi}) + \Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \,\middle|\, \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) \Pr(\widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}) \\
\leq&\Pr(\widehat{\boldsymbol{\psi}} \notin \boldsymbol{\Psi}) + \Pr\left(|V_{ij}^{(k)}(\widehat{\boldsymbol{\psi}}) - V_{ij}(\boldsymbol{\psi}^0)| \geq \epsilon \middle| \widehat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}\right) \\
<&\delta,
\end{aligned}$$

where the last inequality follows from (12) and (13). This proves (9) and thus (8) holds.

Once having (8), the remaining steps to prove Theorem 1 are straightforward. From the second part of (7) of Lemma 1, we have, as $k \to \infty$,

$$\sqrt{k}(\boldsymbol{B}\widehat{\boldsymbol{\beta}} - \boldsymbol{B}\boldsymbol{\beta}^0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{B}\boldsymbol{V}(\boldsymbol{\psi}^0)\boldsymbol{B}^\intercal). \tag{14}$$

By applying the continuous mapping theorem on (8), we have as $k \to \infty$,

$$[\boldsymbol{B}\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}})\boldsymbol{B}^\intercal]^{-\frac{1}{2}} \xrightarrow{p} [\boldsymbol{B}\boldsymbol{V}(\boldsymbol{\psi}^0)\boldsymbol{B}^\intercal]^{-\frac{1}{2}}. \tag{15}$$

Hence, by (14) and (15) and Slutsky's theorem, Theorem 1 is proved. □

Theorem 1 gives the asymptotic distribution of $\boldsymbol{B}\widehat{\boldsymbol{\beta}}$ for any $\boldsymbol{B} \in \mathbb{R}^{n \times 2}$ of rank n. Specifically, if $\boldsymbol{B}$ is the 2×2 identity matrix, then (6) is reduced to the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$. If $\boldsymbol{B} = (0,1)$, then (6) is reduced to the asymptotic distribution of $\widehat{\beta}_1$, which will be used to construct a test statistic for the proposed hypothesis test. We state this specific case in the following corollary.

Corollary 1. *Under Assumptions 1 - 3,*

$$\sqrt{k}[\boldsymbol{V}_{22}^{(k)}(\widehat{\boldsymbol{\psi}})]^{-\frac{1}{2}}(\widehat{\beta}_1 - \beta_1^0) \xrightarrow{d} \mathcal{N}(0,1),$$

as $k \to \infty$*. In addition,* $Z^{(k)} \xrightarrow{d} \mathcal{N}(0,1)$ *as* $k \to \infty$ *under the null hypothesis* H_0*, where* $Z^{(k)} \coloneqq \widehat{\beta}_1[k/\boldsymbol{V}_{22}^{(k)}(\widehat{\boldsymbol{\psi}})]^{1/2}$*.*

Set $Z^{(k)}$ to be the statistic of the hypothesis test. Corollary 1 then implies that the hypothesis test is a Z-test since $Z^{(k)}$ is asymptotically normal. The p-value of the hypothesis test asymptotically equals $2\Phi(-|Z^{(k)}|)$, where Φ denotes the cumulative distribution function of $\mathcal{N}(0,1)$. Moreover, H_0 is rejected at the asymptotic significance level α if $|Z^{(k)}| \geq z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile of $\mathcal{N}(0,1)$, or if the p-value is smaller than α.

Recall that one can easily extend the SESK to incorporate multiple stylized models. We therefore extend the hypothesis test and the related results accordingly as follows. The proof is similar as that of Theorem 1 and Corollary 1 and is omitted.

Theorem 2. *Suppose that* ℓ *distinctive stylized models are used in SESK* (2)*, i.e.,* $\boldsymbol{f}(\boldsymbol{x}) = (1, q_1(\boldsymbol{x}), \ldots, q_\ell(\boldsymbol{x}))^\intercal$ *and* $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_\ell)^\intercal$*. Let* $\boldsymbol{V}^{(k)}(\boldsymbol{\psi}) \coloneqq k(\boldsymbol{F}^\intercal \boldsymbol{\Omega}(\boldsymbol{\psi})^{-1}\boldsymbol{F})^{-1} \in \mathbb{R}^{(1+\ell)\times(1+\ell)}$ *for any* $\boldsymbol{\psi} \in \boldsymbol{\Psi}$*, where* $\boldsymbol{F} = (\boldsymbol{1}, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_\ell) \in \mathbb{R}^{k\times(1+\ell)}$*. Let* $\boldsymbol{B} = (0, 1, \ldots, 1) \in \mathbb{R}^{1\times(1+\ell)}$*. Consider the hypothesis test*

$$H_0 : \beta_1 = \beta_1 = \ldots = \beta_\ell = 0 \quad \textit{versus} \quad H_1 : \beta_1 \neq 0 \textit{ or } \beta_2 \neq 0 \textit{ or } \ldots \textit{ or } \beta_\ell \neq 0.$$

Let $Z^{(k)} \coloneqq [k/(\boldsymbol{B}\boldsymbol{V}^{(k)}(\widehat{\boldsymbol{\psi}})\boldsymbol{B}^\intercal)]^{1/2} \cdot \boldsymbol{B}\widehat{\boldsymbol{\beta}}$ *be the test statistic. Then,* $Z^{(k)} \xrightarrow{d} \mathcal{N}(0,1)$ *as* $k \to \infty$*, under Assumptions 1 - 3 and the null hypothesis* H_0*.*

Remark 4. Rejecting the null hypothesis in the Z-test only means that at least one of the stylized models included is useful in capturing the shape of the response surface. It does not, however, indicate which stylized model is the best or which combination of these stylized models should be chosen. To quantify the effectiveness of the stylized models included in SESK, one can use the K^2 statistic presented in §3.2 or the information criteria like AIC and BIC presented in §3.3.

3.2 K^2 Statistic for Effectiveness

In the presence of multiple candidate stylized models, it is desirable for the users to have a convenient tool for model selection. In the context of linear regression, the coefficient of determination R^2 measures how well a model fits the data and can be interpreted as the proportion of the total variation of the data explained by the model. However, R^2 or other measures based on the sum of squared errors may not be suitable in our setting for the following reason. For stochastic kriging, in the absence of simulation noise, the predicted surface would pass through exactly the simulation outputs at the design points, in which case the sum of squared errors would be zero. This implies that the sum of squared errors can be reduced to zero by simply increasing the number of replications

at each design point, and thus cannot reflect the goodness-of-fit of the stylized model. Nevertheless, inspired by R^2, we propose a new statistic called K^2:

$$K^2 = 1 - \frac{\widehat{\tau}_S^2}{\widehat{\tau}_O^2},$$

where $\widehat{\tau}_S^2$ and $\widehat{\tau}_O^2$ are the ML estimates of τ^2 in SESK and OSK, respectively.

Similar as R^2, a large value of K^2 indicates that a large portion of the variation in the observations $\bar{\boldsymbol{Y}}$ can be explained by the stylized model. An intuitive reason is as follows. OSK treats the response surface $\mathcal{Y}(\boldsymbol{x})$ as a realization of a second-order stationary Gaussian process, which has a *constant* marginal variance. In order to capture the highly nonlinear shape of the response surface for queueing simulation, which often exhibits exploding behavior, the marginal variance τ_O^2 in OSK needs to be large. By contrast, the same process is used to model the *residual* surface $\mathcal{Y}(\boldsymbol{x}) - \beta_1 q(\boldsymbol{x})$ in SESK. If the stylized model $q(\boldsymbol{x})$ can capture the main trend of $\mathcal{Y}(\boldsymbol{x})$, the residual surface would have much less variation than the original surface does, and thus the marginal variance τ_S^2 in SESK would be smaller than τ_O^2.

Remark 5. Consider the following two linear regression models that are analogous to OSK and SESK, respectively: (i) $y_i = c_O + \sigma_O \epsilon_i$ and (ii) $y_i = c_S + \beta x_i + \sigma_S \epsilon_i$, $i = 1, \ldots, k$, where ϵ_i's are independent standard normal random variables. Let $\widehat{\sigma}_O$, $\widehat{\sigma}_S$, $\widehat{c}_S$, and $\widehat{\beta}$ denote the ML estimates. It can be shown easily that $\widehat{\sigma}_O^2 = k^{-1} \sum_{i=1}^k (y_i - \overline{y})^2$, where $\overline{y} = k^{-1} \sum_{i=1}^k y_i$, and $\widehat{\sigma}_S^2 = k^{-1} \sum_{i=1}^k (y_i - \widehat{c}_S - \widehat{\beta} x_i)^2$. Hence, R^2 associated with the simple linear regression is

$$R^2 = 1 - \frac{\text{sum of squared residuals}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^k (y_i - \widehat{c}_S - \widehat{\beta} x_i)^2}{\sum_{i=1}^k (y_i - \overline{y})^2} = 1 - \frac{\widehat{\sigma}_S^2}{\widehat{\sigma}_O^2},$$

which bears a structure similar to K^2.

Remark 6. It is well known that the value range of R^2 is $[0, 1]$, but K^2 is more subtle. We prove in Appendix A that $K^2 \in [0, 1]$ if $\boldsymbol{\Sigma} = \boldsymbol{0}$ and $\boldsymbol{\theta}$ is known. Otherwise, K^2 may become negative, albeit rarely in practice, when the number of design points k is small. A negative K^2 means complete failure of the stylized model.

We now provide another intuition regarding K^2. Let $\text{IMSE} = \int \text{MSE}^*(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ denote the integrated MSE of stochastic kriging over the entire design space, where $\text{MSE}^*(\boldsymbol{x})$ is given by (3). Then, $\frac{\text{IMSE}_S}{\text{IMSE}_O}$ can also be used to measure the global goodness-of-fit of SESK relative to OSK, where the subscript S and O indicate SESK and OSK, respectively. But it may be computationally prohibitive to implement this measure in practice because it involves a multidimensional numerical integration and, even to approximate it requires computation of the true responses at a large number of design points, which would be excessively expensive.

Nevertheless, we argue heuristically that $1 - \frac{\text{IMSE}_S}{\text{IMSE}_O}$ is somewhat similar to K^2 intuitively. Assume that the numbers of replications at all design points are sufficiently large so that the simulation errors there are negligible. Then, it can be shown easily that $\text{MSE}^*(\boldsymbol{x}_0) \approx 0$ if $\boldsymbol{x}_0$ is close to any one

of the design points and that $\mathrm{MSE}^*(\boldsymbol{x}_0) \approx \tau^2$ if $\boldsymbol{x}_0$ is far away from all the design points. Therefore, if the design space is large and the number of design points is small, then IMSE is roughly proportional to τ^2, and thus $K^2 = 1 - \frac{\tau_S^2}{\tau_O^2} \approx 1 - \frac{\mathrm{IMSE}_S}{\mathrm{IMSE}_O}$.

We stress here that albeit inspired by R^2, K^2 is not rigorously derived and should only be considered as a *heuristic* statistic that measures the proportion of the total variation of the response surface explained by the incorporated stylized model. But its advantage is the simplicity, since the estimation of τ^2 is necessary for use of stochastic kriging and K^2 can be computed with nearly zero additional cost. Consequently, we suggest that K^2 should not be used alone but as a sanity check to ensure that conclusions from other statistical tools such as the Z-test in §3.1 and the information criteria in §3.3 are consistent.

3.3 Information Criteria

Although we have been focusing on the case of one single stylized model in SESK, multiple stylized models can indeed be incorporated. Then, the complexity of SESK metamodels may be different in terms of the number of unknown parameters. So in addition to K^2, which does not account for the metamodel complexity, we suggest using the popular model selection methods Akaike information criterion (AIC, Akaike 1974) and Bayesian information criterion (BIC, Schwarz 1978). Both criteria are driven by MLE and penalize the number of model parameters in an effort to avoid model overfitting. Their difference lies in the form of the penalization. In particular,

$$\begin{aligned}\mathrm{AIC} &= -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}}) + 2p,\\ \mathrm{BIC} &= -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}}) + p\ln(k),\end{aligned}$$

where p is the number of unknown parameters and k is the data size (i.e. number of design points). For example, $p = m+1$ for OSK since $\boldsymbol{\beta} = \beta_0$, where m is the size of $(\tau^2, \boldsymbol{\theta})$, whereas $p = m+\ell+1$ for SESK, where ℓ is the number of stylized models incorporated. If several SESK metamodels are available, we select the one with the smallest AIC or BIC value. Notice that AIC or BIC can also be used to select the best among several stylized models if one wants to incorporate one single stylized model in SESK.

4 An Illustrative Example – $M/G/1$ Queue

We now consider a simple example to gain insights on benefits of incorporating stylized models in stochastic kriging, and demonstrate the proposed measures. Let $\mathcal{Y}(x)$ be the steady-state mean queue length (excluding the customer in service) of an $M/G/1$ queue with arrival rate $x \in (0,1)$ and unit service rate. Suppose that the service times follow the gamma distribution with shape parameter $1/2$ and scale parameter 2, so the squared coefficient of variation of the service time distribution is 2. It then follows from the Pollaczek-Khintchine formula that $\mathcal{Y}(x) = 1.5x^2/(1-x)$.

Let $Q_t(x)$ be the queue length of this system at time t. A natural estimator of $\mathcal{Y}(x)$ is

$$\bar{Y}_T(x) \coloneqq \frac{1}{T}\int_0^T Q_t(x)\,\mathrm{d}t,$$

the average queue length during T units of simulated time, and its *asymptotic variance* $\sigma^2(x) \coloneqq \lim_{T\to\infty} T\,\mathrm{Var}[\bar{Y}_T(x)]$ is given by Equation (13) of Whitt (1989)

$$\sigma^2(x) = \frac{x(20+121x-116x^2+29x^3)}{4(1-x)^4}.$$

Then, for large T, $\bar{Y}_T(x) \overset{d}{\approx} \mathcal{N}(\mathcal{Y}(x), \sigma^2(x)/T)$, where $\overset{d}{\approx}$ means "approximately equals in distribution". Therefore, we can use this approximation to generate random samples of $\bar{Y}_T(x)$ instead of running steady-state simulation, which is time-consuming and subject to the initialization bias.

4.1 Benefits of Stylized Models and Validity of Measures

To illustrate SESK, we set up an experiment with $k = 4$ design points $x = 0.2, 0.4, 0.6, 0.8$, and allocate $n = 20$ simulation replications to each of them. Each replication is generated from $\mathcal{N}(\mathcal{Y}(x), \sigma^2(x)/T)$ with $T = 2,500$. Based on the synthetic data, we compute the estimates $(\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\beta}}, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$, the plug-in predictor $\widehat{\widehat{\mathcal{Y}}}(x)$, and the MSE estimator $\widehat{\mathrm{MSE}}(x)$.

Consider three stylized models: $q^{(1)}(x) = x^2/(1-x)$, $q^{(2)}(x) = 3x^9$, and $q^{(3)}(x) = 10(x-0.52)^2$. Notice that $q^{(1)}(x)$ is the steady-state mean queue length in an $M/M/1$ queue with arrival rate x and unit service rate, which is obviously a good approximation for the $M/G/1$ queue. As a rough approximation, $q^{(2)}(x)$ somewhat captures the trend of $\mathcal{Y}(x)$, albeit not as closely as $q^{(1)}(x)$ does. Last, $q^{(3)}(x)$ appears irrelevant to $\mathcal{Y}(x)$. For instance, $q^{(3)}(x)$ is not an increasing function as $\mathcal{Y}(x)$. Notice that all these stylized models are in closed form, so the computational cost is negligible compared to running the simulation model.

Figure 2 shows that OSK completely fails to capture the exploding behavior of $\mathcal{Y}(x)$ as x approaches 1. By contrast, incorporating the response surface of the $M/M/1$ queue $q^{(1)}(x)$ yields a predicted surface almost identical to the true surface. Of course, this is an ideal case since $q^{(1)}(x)$ happens to be a multiple of $\mathcal{Y}(x)$. A more realistic situation is demonstrated by $q^{(2)}(x)$, which captures the monotonicity but not the exploding behavior of $\mathcal{Y}(x)$. Surprisingly, although $q^{(2)}(x)$ is merely a rough approximation, it dramatically enhances the prediction accuracy of SESK relative to OSK. This shows that SESK is fairly robust with respect to the choice of the stylized model. Nonetheless, if the incorporated stylized model has little similarity to the true surface, which is the case for $q^{(3)}(x)$, SESK does not show significant improvement over OSK.

Table 1 confirms the above findings. The Z-test suggests that $q^{(1)}(x)$ and $q^{(2)}(x)$ are significant for explaining the variation in $\mathcal{Y}(x)$, whereas $q^{(3)}(x)$ is not. Moreover, the three statistics K^2, AIC, and BIC all indicate that $q^{(1)}$ and $q^{(3)}(x)$ are the best and worst among the three stylized models, respectively. In particular, $K^2 \approx 1$ for $q^{(1)}(x)$ aligns with the fact that $q^{(1)}(x)$ perfectly captures

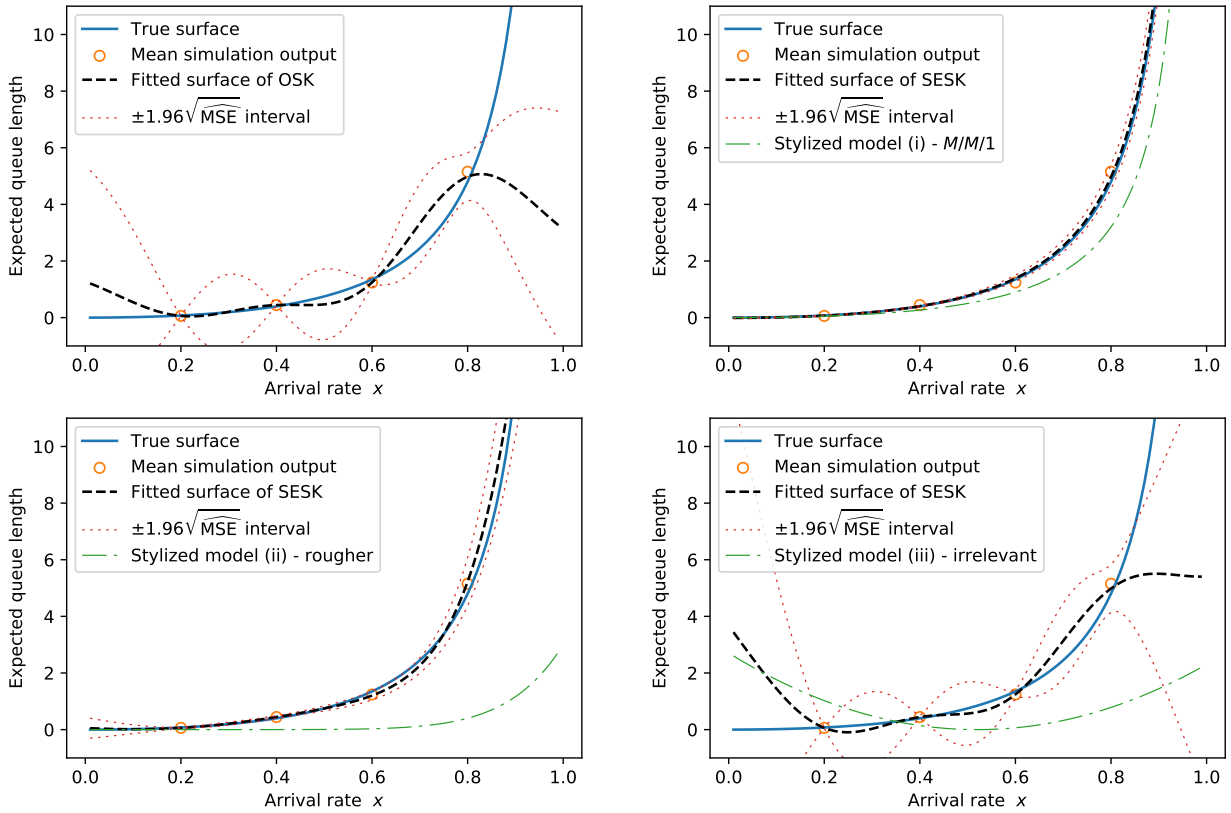

Figure 2: OSK v.s. SESK for the $M/G/1$ Queue.

the trend of $\mathcal{Y}(x)$, making the variation of the residual surface $\mathcal{Y}(x) - \beta_1 q^{(1)}(x)$ negligible.

Table 1: OSK v.s. SESK for the $M/G/1$ Queue.

Metamodel	Z-test		K^2	AIC	BIC
	$Z^{(k)}$	p-value			
OSK	-	-	-	22.7	20.9
SESK with $q^{(1)}$	22.4	<0.001	1.00	-1.3	-3.8
SESK with $q^{(2)}$	7.6	<0.001	0.95	9.9	7.4
SESK with $q^{(3)}$	0.6	0.54	0.06	24.4	21.9

Albeit artificial, this numerical example demonstrates that by incorporating a reasonable stylized model, SESK can produce substantially more accurate predictions than OSK, even with only a small number of design points. More importantly, SESK is relatively robust and does not require the incorporated stylized model to have a high degree of similarity to the true response surface. Finally, the proposed Z-test and K^2 statistic can diagnose the usefulness and effectiveness of a candidate stylized model for the small sample case, producing findings that are consistent with the popular tools AIC and BIC.

4.2 Comparison with Gradient Enhanced Stochastic Kriging

Another approach to enhancing stochastic kriging is to incorporate gradient information. We now compare SESK with the stochastic kriging with gradient (SKG) proposed in Chen et al. (2013) in terms of prediction accuracy; see Qu and Fu (2014) for a different way of incorporating gradient estimates in stochastic kriging that has comparable performance with SKG. However, in order that gradient enhanced stochastic kriging be feasible, one needs to be able to compute the gradient estimates with a negligible additional cost once the responses of the simulation model have been observed. To that end, the simulation model ought to have a relatively simple structure, so that techniques such as infinitesimal perturbation analysis or likelihood ratio method can be applied to compute the gradient estimates efficiently. However, the general simulation models considered in this paper are too complex to yield efficient gradient estimation; see examples in §6. One then often resorts to the finite difference (FD) method, which amounts to a substantial computational overhead by running the simulation model at nearby locations and tends to incur significant estimation errors.

We conduct the comparison between SESK and SKG for the $M/G/1$ queue. To imitate the usual situation in practice where the simulation model is a black-box and the user does not have a means to compute the gradients efficiently, we apply FD to estimate the gradients. In particular, associated with each simulation replication at each design point, we compute a FD gradient estimate based on the central difference with step size 0.01. The results are presented in Figure 3. The left panel shows the predicted surface of SKG, whereas the right panel shows the predicted surface of SESK with a coarse stylized model $q^{(2)}$. Clearly, SESK has better performance in terms of extrapolating the exploding response surface. This is because the gradient estimates only provide local information about the shape of the surface, whereas the stylized model provides global information. In addition, using central differences for gradient estimation, SKG would run the simulation model 3 times as many as SESK. Therefore, if the stylized model is well chosen, then SESK can produce more accurate prediction with less computational cost than SKG does.

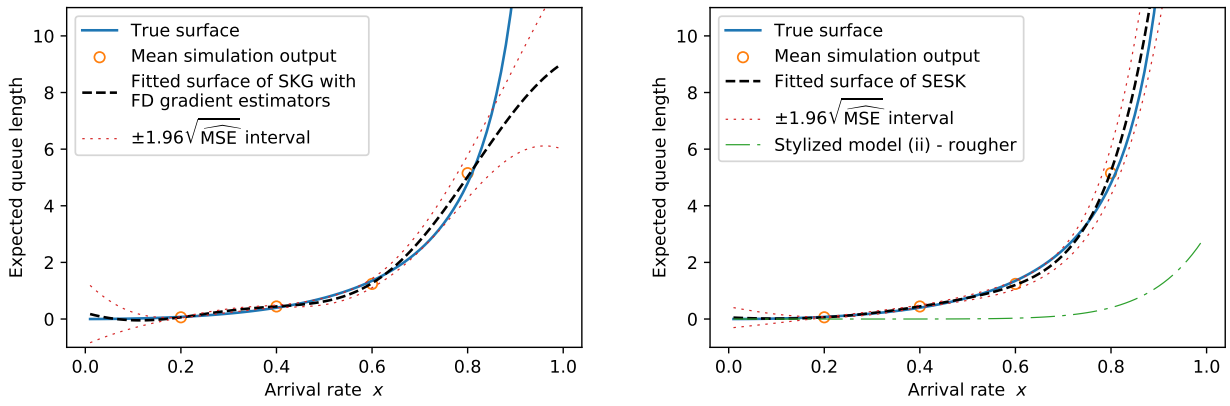

Figure 3: SKG v.s. SESK for the $M/G/1$ Queue.

5 Constructing Stylized Models for Queueing Networks

Clearly, there is no unique rule for constructing stylized models for queueing simulation. A stylized model with higher accuracy is generally more complex and more computationally expensive. We remark here that the *quantitative* accuracy of the stylized model is not our uppermost concern. Instead, we focus on the simplicity and convenience of the approaches, since a basic representation of the *qualitative* behavior of the response surface may suffice.

Suppose that the simulation model is an open queueing network that consists of a finite number of stations to provide service. External customers may enter the network via each station. A customer that completes the service at one station may be routed to another to receive further service or leave the network. In this section, we propose several simple methods for constructing stylized models for such queueing networks. The resulting stylized models either yield closed-form solutions or can be computed numerically with a negligible cost relative to the simulation model.

5.1 Jackson Network

The Jackson network (Jackson 1963) is a classical stylized queueing model. It assumes that the external arrivals to each station follow a Poisson process, that the service times at each station are independent and exponentially distributed, and that each station has an infinite capacity. Moreover, the customers are routed randomly according to prespecified probabilities. The Jackson network is highly analytically tractable. Many performance measures such mean waiting time and mean number-in-system have closed-form expressions. Analogous to the $M/G/1$ queue in §4, one may consider to use the Jackson network as the stylized model in SESK if each station of the original queueing network has stationary external arrivals, independent and identically distributed service times, and infinite capacity.

5.2 Finite Capacity and Blocking

Service capacity is often a design variable that needs to be optimized in practice to balance system performance and operating costs. Hence, networks with infinite capacity like the Jackson network are not suitable to approximate such real systems. Notice that finite capacity queueing networks have potential loss and temporary blocking of customers; see Balsamo et al. (2013, §2.2).

In queueing theory, there is a class of analytically tractable queueing networks that permit the so-called "product-form" solutions. Examples include Jackson networks (Jackson 1963) and Kelly networks (Kelly 1979); see Chao (2011) for an recent overview. Such networks can be decomposed into isolated stations and their behavior can be analyzed separately. This suggests that if the simulation model is a finite-capacity queueing network, then one may simply consider the decomposition approximation as the stylized model for SESK. The only nontrivial calculation required is to properly adjust the parameters such as arrival rates and service rates to approximate the interdependence between the stations in the original network. We provide in Appendix B a relatively simple way for parameter adjustment based on the approach in Korporaal et al. (2000).

The case study in §6.1 demonstrates the use of the stylized model constructed in this way to study patient flow in a hospital. We refer interested readers to Osorio and Bierlaire (2009) for a more sophisticated approximation scheme.

5.3 Time-Varying Arrivals

The arrival process is often nonstationary in practice. A natural approach to addressing time-varying arrivals is the pointwise stationary approximation (Whitt 1991). In this approach, the time-varying arrival rate process is approximated by a piecewise constant function. Within each piece, the arrival rates are considered as a constant equal to the average. The performance measure of interest can be calculated independently for each piece and then aggregated by taking a weighted average, where the weights may be approximated by the total number of customers in each piece. The advantage of this approach is its simplicity. However, the performance of this approach depends critically on how the pieces are set in terms of both the number of pieces and the length of each piece. In addition, during the peak time of the arrivals, the average arrival rate within a piece may exceed the total service rate, implying an unstable queue.

Another approach is the fluid approximation that is built on heavy traffic analysis; see Gautam (2012, §8) for an introduction on the subject. We provide in Appendix C a simple fluid approximation for time-varying queues. Both of the above approaches for constructing stylized models for time-varying arrivals will be demonstrated in the case study in §6.2 that addresses the dock allocation problem at an air cargo terminal.

6 Case Studies

We consider two real-world examples of queueing simulation in this section, each of which has a distinctive feature. The first example stems from the healthcare industry and involves a queueing network with finite capacity in each of its stations which induces blocking behavior. The second example, on the other hand, comes from the logistics industry and involves time-varying arrival processes. We demonstrate, through the two case studies, that: (i) reasonable stylized models can be constructed easily for a large class of queueing networks; (ii) prediction accuracy of stochastic kriging can be significantly enhanced by incorporating such stylized models; (iii) the proposed measures can diagnose and quantify the improvement properly. All the numerical experiments (including the $M/G/1$ queue in §4) are implemented in MATLAB R2015a (Intel i7-3770 CPU @ 3.40GHz, 8 GB RAM). The code is available at `simopt.github.io`. In the implementation, we have taken advantage of the open source code for stochastic kriging (`stochastickriging.net`).

6.1 Case Study 1 – Patient Flow in a Hospital

This problem is adopted from Osorio and Bierlaire (2009). The hospital of interest has nine medical units (i.e., stations), each of which has a different number of beds. The patients and the beds are considered as customers and servers, respectively. For medical unit i, $i = 1, \ldots, 9$, the external

patients arrive following a Poisson process with arrival rate γ_i, the service time of each bed follows the exponential distribution with rate μ_i, and the number of beds is c_i; see Table 2. The hospital as a whole is modeled as a queueing network and the patients are routed among the medical units according to the transition probability matrix given in Table 3. Each medical unit is bufferless having no waiting room, so the capacity of each unit is the same as the number of beds. Due to the finite capacity, a patient who finishes his service in one unit will be blocked at his current location, if the unit to which he should be routed is full. In this case, the patient waits at his current location until there is an opening in the target unit. While being blocked, the patient keeps occupying his bed and it is unavailable for other patients. If there are multiple blocked patients waiting to enter the same unit, they are unblocked on a first-blocked-first-released basis. This blocking mechanism is known as blocking-after-service (BAS) (Balsamo et al. 2013, §2.2). The performance measure of interest $\mathcal{Y}(\boldsymbol{x})$ is the steady-state sojourn time in the hospital (length of stay), where $\boldsymbol{x} \in \mathbb{R}^9$ is the vector of the number of beds in the nine medical units.

Table 2: Parameter Configuration of the Hospital.

i	1	2	3	4	5	6	7	8	9
γ_i/hr.	0.39	0.50	0.25	0.06	0.18	0.03	0.13	0.16	0.00
μ_i/hr.	0.32	0.26	0.34	0.01	0.02	0.01	0.02	0.22	0.52
c_i	4	8	5	18	18	4	4	10	6

Table 3: Transition Probability Matrix.

i	1	2	3	4	5	6	7	8	9
1	-	-	-	.16	.02	-	-	.71	-
2	-	-	-	.07	-	-	-	.84	-
3	-	-	-	.03	.01	-	-	-	.95
4	.18	.01	.03	-	.03	.01	.11	.03	-
5	.05	.01	.01	.01	-	.07	-	-	-
6	.02	-	-	.01	.10	-	-	-	-
7	.05	-	.05	.04	-	-	-	.01	-
8	-	-	-	-	-	-	.01	-	-
9	-	-	-	.05	-	-	.05	.02	-

We want to predict $\mathcal{Y}(\boldsymbol{x})$ over a wide range of $\boldsymbol{x}$, say, up to 60% difference from the current configuration, i.e. $x_i \in (1 \pm 60\%)c_i$, $i = 1, \ldots, 9$. To implement the metamodels, we consider a simple experiment design. We alter the value of each x_i while keeping the other variables fixed at the current configuration, namely the design point is of the form $(c_1, \ldots, c_{i-1}, x_i, c_{i+1}, \ldots, c_9)$. The altered values are chosen to be centered around the current configuration of the number of beds in each station; see the row of "Design Point" in Table 4. Including these altered configuration and the current configuration of the number of beds, we obtain 23 design points in total and use them to construct the metamodels. At each design point, the simulation model is run with 10 replications with warm-up length 10,000 hours and total run length 50,000 hours. The average run time of the

simulation model at one design point is about 240 seconds.

Table 4: Experimental Design. The numbers in the table are the possible values of x_i. The design points given here do not but should include the current configuration $(c_1, \ldots, c_9)$.

i	1	2	3	4	5	6	7	8	9
Design Point $(c_1, \ldots, c_{i-1}, x_i, c_{i+1}, \ldots, c_9)$	2, 6	4, 12	2, 8	10, 14 22, 26	10, 14, 22, 26	2, 6	2, 6	6, 14	2, 10
Evaluation Point $(x_1, \ldots, x_9)$	3, 5	3, 13	3, 7	8, 18, 28	8, 18, 28	3, 5	3, 5	5, 15	3, 9

To apply SESK, we construct a simple stylized model as follows. We decompose the finite-capacity queueing network into isolated independent stations and model each station as a finite-capacity multi-server queue. The only non-trivial calculation required is to adjust the structural parameters of each station (e.g., arrival and service rates) properly in order to approximate the interdependence between the stations in the original network. The whole construction is based on the approach of Korporaal et al. (2000), and the details are given in Appendix B. The average run time of evaluating the stylized model at one design point is about 0.06 second, thereby negligible compared to the simulation model.

To evaluate the prediction accuracy, we consider the root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\sum_{\boldsymbol{x}_0 \in \mathcal{C}} \frac{1}{|\mathcal{C}|} (\widehat{\widehat{\mathcal{Y}}}(\boldsymbol{x}_0) - \mathcal{Y}(\boldsymbol{x}_0))^2}, \tag{16}$$

and the mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{100}{|\mathcal{C}|} \sum_{\boldsymbol{x}_0 \in \mathcal{C}} \left| \frac{\widehat{\widehat{\mathcal{Y}}}(\boldsymbol{x}_0) - \mathcal{Y}(\boldsymbol{x}_0)}{\mathcal{Y}(\boldsymbol{x}_0)} \right|, \tag{17}$$

where $\mathcal{C}$ denotes the set of predicted points and $|\mathcal{C}|$ is its cardinality. However, the total number of points in the entire domain is large and to estimate the "true" response at any point requires extensive simulation. It is thus computationally prohibitive to use all these points to evaluate the prediction quality of the metamodels. Instead, we select several representative values for each x_i, and use the full factorial design to form the grid for the evaluation purpose. We call these points the evaluation points; see the row of "Evaluation Point" in Table 4. The total number of evaluation points is $|\mathcal{C}| = \prod_{i=1}^{9} m_i = 1,152$, where m_i is the number of representative values of x_i. For each of the 1,152 evaluation points, we estimate the "true" response through running the simulation for sufficiently many replications so that the simulation noise is negligible (e.g. the half-width of the 95% confidence interval is less than 0.05 hour).

The numerical results are shown in Table 5. We can find that both the RMSE and MAPE in SESK are smaller than those in OSK, which shows that the SESK gives better fitting. On the other hand, the results of the hypothesis test strongly reject H_0, suggesting the usefulness of the stylized

model. The K^2 is 0.82, showing high explaining power of the stylized model. Besides, the AIC and BIC in SESK are much smaller than those in OSK. All the measures are consistent to the actual improved performance by incorporating the stylized model.

Table 5: OSK v.s. SESK for Patients' Mean Sojourn Time.

Metamodel	Hypothesis test		K^2	AIC	BIC	RMSE (hr.)	MAPE (%)
	$Z^{(k)}$	p-value					
OSK	-	-	-	49.8	62.3	3.7	7.4
SESK	12.9	< 0.001	0.82	15.7	29.3	1.3	2.7
Stylized Model	-	-	-	-	-	2.3	5.4

For better visualization, we plot in Figure 4 the predicted patients' mean sojourn time as a function of x_4 while keeping the other variables fixed, i.e. $\mathcal{Y}(\boldsymbol{x})$ for $\boldsymbol{x} \in \{\boldsymbol{x} | \boldsymbol{x} = (4, 8, 5, x_4, 14, 4, 4, 10, 6)^\intercal\}$. Notice that in this set, the point with $x_4 = 18$ happens to be a design point at which the simulation model is executed. We see clearly that the prediction accuracy of OSK is satisfying for the flat part of the response surface, which is expected. However, it does not capture the steep, nonlinear part at the left end of the curve. This is the place where the stylized model makes an impact and improves the prediction accuracy considerably.

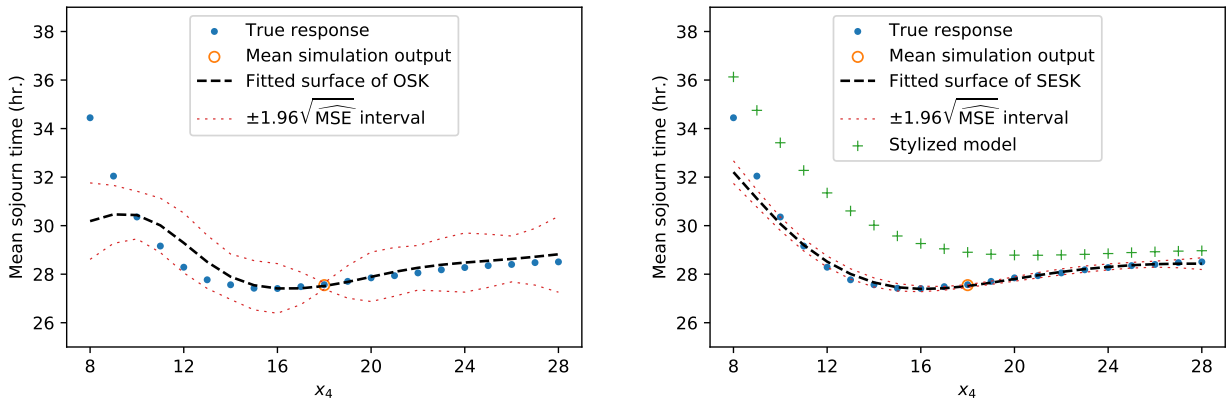

Figure 4: Predicted Patients' Mean Sojourn Time as a Function of x_4.

6.2 Case Study 2 – Dock Allocation at an Air Cargo Terminal

This is a practical problem encountered by one of the largest air cargo terminals worldwide. A critical resource for daily operations of the terminal is shipping/receiving docks, where cargoes are delivered and picked up at the docks by trucks of forwarding agents. There are primarily four types of cargo at this terminal: (1) pallet bulk cargo, (2) general bulk cargo, (3) perishable cargo, and (4) prepacked cargo. Each type of cargo demands a distinctive material handling system that does not apply to other types. Hence, the management of the terminal needs to determine the optimal scheme for allocating the total available docks to the four types of cargo so that the average waiting

time of the trucks is minimized.

The terminal operates continuously (24 hours per day, 7 days per week). The historical data indicate that the arrival rates of the four types of cargoes are time-varying (Figure 5) and that their service times follow different probability distributions (Table 6). The simulation model of interest here is the terminal consisting of four independent $M_t/G/s$ queues, each of which models the process of handling one type of cargo with s allocated docks. Let $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\intercal$ be the numbers of docks that are allocated to the four types of cargoes. The performance measure of interest, $\mathcal{Y}(\boldsymbol{x})$, is the long-run average waiting time of the trucks.

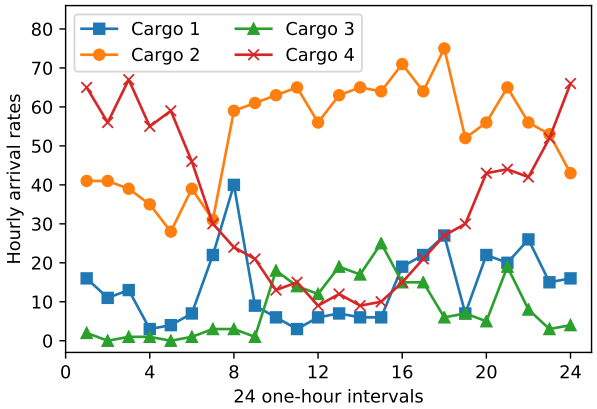

Figure 5: Time-Varying Arrival Rates of the Four Cargo Types.

Table 6: Service Time Distributions and Dock Allocation. WEIB(a, b) and GAMM(a, b) denote the Weibull and Gamma distribution with scale parameter a and shape parameter b, respectively.

Cargo Type	Service Time Distribution (min.)	Number of Docks
1	WEIB(21.8, 1.3)	x_1
2	7 + WEIB(67.6, 1.5)	x_2
3	7 + GAMM(25.7, 0.9)	x_3
4	7 + GAMM(9.4, 3.0)	x_4

We consider two stylized models. One (i.e., stylized model 1) is the stationary approximation, that is, queue i is approximated by an $M/M/s_i$ queue with arrival rate $\bar{\lambda}_i$ and service rate μ_i, where $\bar{\lambda}_i$ is the average arrival rate and $s_i = x_i$, $i = 1, \ldots, 4$. The other (i.e., stylized model 2) is based on the fluid approximation of time-varying queues and is detailed in Appendix C. The average run time of evaluating the two stylized models at one design point is about 0.002 and 0.003 second, respectively.

6.2.1 Evaluating Prediction Accuracy

The terminal has 111 available docks in total, so $x_1 + x_2 + x_3 + x_4 \leq 111$. To ensure stability of the four queues, the number of docks must satisfy $\bar{\lambda}_i < \mu_i x_i$, $i = 1, \ldots, 4$. Specifically, this requirement translates to $\{x_1 \geq 5, x_2 \geq 61, x_3 \geq 5, x_4 \geq 21\}$. To evaluate the prediction accuracy, we calculate

RMSE defined in (16) and MAPE defined in (17), for all points in the feasible region. Hence, $|\mathcal{C}| = 8,855$. For each evaluation point, we estimate the "true" response by extensive simulation as follows. The run length of the simulation model is set sufficiently long (e.g. 25 days = 600 hours) and the first 10 days is considered as the warm-up period to in order to reduce the initialization bias. Moreover, each simulation is replicated for sufficiently many times so that the simulation noise is small (e.g. the half-width of the 95% confidence interval is less than 0.1 minute).

Since the feasible region is not a hypercube, to find a good set of design points we adopt the space-filling approach detailed in Forrester et al. (2008, §1.4.3), while considering uniform random designs instead of random Latin hypercubes. Given the total number of design points, the optimal design is defined via maximizing the smallest pairwise distance between the design points (Morris and Mitchell 1995). However, to identify the optimal design is computationally infeasible in general, so certain random search procedures are often used to find the best design up to a computational budget. We assume that the number of design points is 40 and repeat the above space-filling approach 40 times, each using a different random seed and resulting in a different design upon termination. We conduct the experiment for each of the 40 designs. To produce the simulation outputs for constructing the metamodels, normally at each design point the simulation model is replicated 30 times with the run length 8 days and warm-up period 5 days. For design points at which the estimated standard deviation of the simulation output is beyond 1 minute, simulation effort is increased so that it is controlled around 1 minute. The average run time of the simulation model at one design point is about 96 seconds. Hence, the computational effort for the two considered stylized models (0.002 and 0.003 second run time) is indeed negligible.

In addition to the single stylized-model setup, we also consider the case of incorporating both stylized models in SESK, in which case the trend term becomes $\beta_0 + \beta_1 q_1(\boldsymbol{x}) + \beta_2 q_2(\boldsymbol{x})$. The numerical results based on one typical set of design points are presented in Table 7, and those based on all the 40 sets of design points are presented in Table 8.

Table 7: OSK v.s. SESK for Trucks' Mean Waiting Time Based on One Typical Design.

Metamodel	Z-test		K^2	AIC	BIC	RMSE (min.)	MAPE (%)
	$Z^{(k)}$	p-value					
OSK	-	-	-	305.3	315.5	9.6	10.7
SESK (stylized model 1)	21.1	<0.001	0.62	260.0	271.9	4.3	4.4
SESK (stylized model 2)	74.2	<0.001	0.86	208.9	220.7	2.3	2.6
SESK (stylized model 1 & 2)	44.1	<0.001	0.99	137.5	151.0	1.0	1.1
Stylized model 1	-	-	-	-	-	44.6	68.7
Stylized model 2	-	-	-	-	-	21.8	22.2

Clearly, the stationary approximation is very rough for queueing systems with time-dependent characteristics. Despite its crudeness, it can significantly improve the prediction accuracy of the stochastic kriging metamodel. The fluid approximation is a much better stylized model, suggested

Table 8: OSK v.s. SESK for Trucks' Mean Waiting Time Based on 40 Designs.

Kriging model	RMSE (min.)				MAPE (%)			
	Min	Max	Mean	Median	Min	Max	Mean	Median
OSK	6.2	15.7	9.8	9.4	7.5	18.6	10.7	10.1
SESK (stylized model 1)	2.2	11.7	3.6	3.1	2.2	7.4	3.4	3.2
SESK (stylized model 2)	1.8	4.8	2.5	2.3	2.1	3.5	2.7	2.6
SESK (stylized model 1 & 2)	0.9	1.4	1.1	1.0	1.1	1.6	1.2	1.2

by both the model selection tools (K^2, AIC, and BIC) and the measurements for prediction accuracy (RMSE and MAPE). Incorporating both stylized models provide even more accurate predictions.

In addition, based on the same design as that of Table 7, we plot in Figure 6 the predicted long-run mean waiting time of different metamodels as a function of x_2, i.e. $\mathcal{Y}(\boldsymbol{x})$ for $\boldsymbol{x} \in \{\boldsymbol{x}|x_1 = 6, x_2 = 62, \ldots, 74, x_3 = 10, x_4 = 95 - x_2\}$. Notice that $(6, 69, 10, 26)$ in the set happens to be a design point at which the simulation model is executed. As expected, OSK does not perform well and stylized models greatly improve the prediction accuracy. In particular, stylized model 1 roughly captures the trend of the response surface and thus provides noticeable improvement; stylized model 2 is more accurate and the improvement it causes is substantial, making the predicted responses almost identical to the true responses. Incorporating both stylized models further improves the prediction accuracy as shown by the better fitted surface and narrower $\pm 1.96\sqrt{\widehat{\text{MSE}}}$ intervals. Notice also that in terms of the prediction MSE, SESK with stylized model 1 incorporated is highly accurate for $x_2 > 67$, whereas SESK with stylized model 2 incorporated is much better for $x_2 \leq 67$. Hence, incorporating both stylized models is, to some extent, analogous to model averaging (Hastie et al. 2009, Chapter 8).

6.2.2 Searching for Optimal Allocation

To find the optimal dock allocation is a discrete optimization via simulation (DOvS) problem. Random search algorithms are often used to solve such problems; see Andradóttir (2006) for an overview. In particular, the Gaussian process-based search (GPS) algorithm developed in Sun et al. (2014) is a state-of-the-art globally convergent algorithm. In the same vein as OSK, the GPS algorithm treats the response surface $\mathcal{Y}(\boldsymbol{x})$ as a realization of a second-order stationary Gaussian process. In each iteration of the algorithm, based on the constructed Gaussian process one can calculate the probability that a solution is better than the current sample-best solution. This probability is then used to build a sampling distribution for the next design point to be sampled. It is shown in Sun et al. (2014) that in terms of total computational time including both running time of the simulation model and computational overhead, the GPS algorithm has significantly better performance than other popular algorithms, such as the global random search algorithm of Andradóttir (1996) and the sequential kriging optimization algorithm of Huang et al. (2006), for typical DOvS problems. We now use the dock allocation problem to illustrate how stylized models

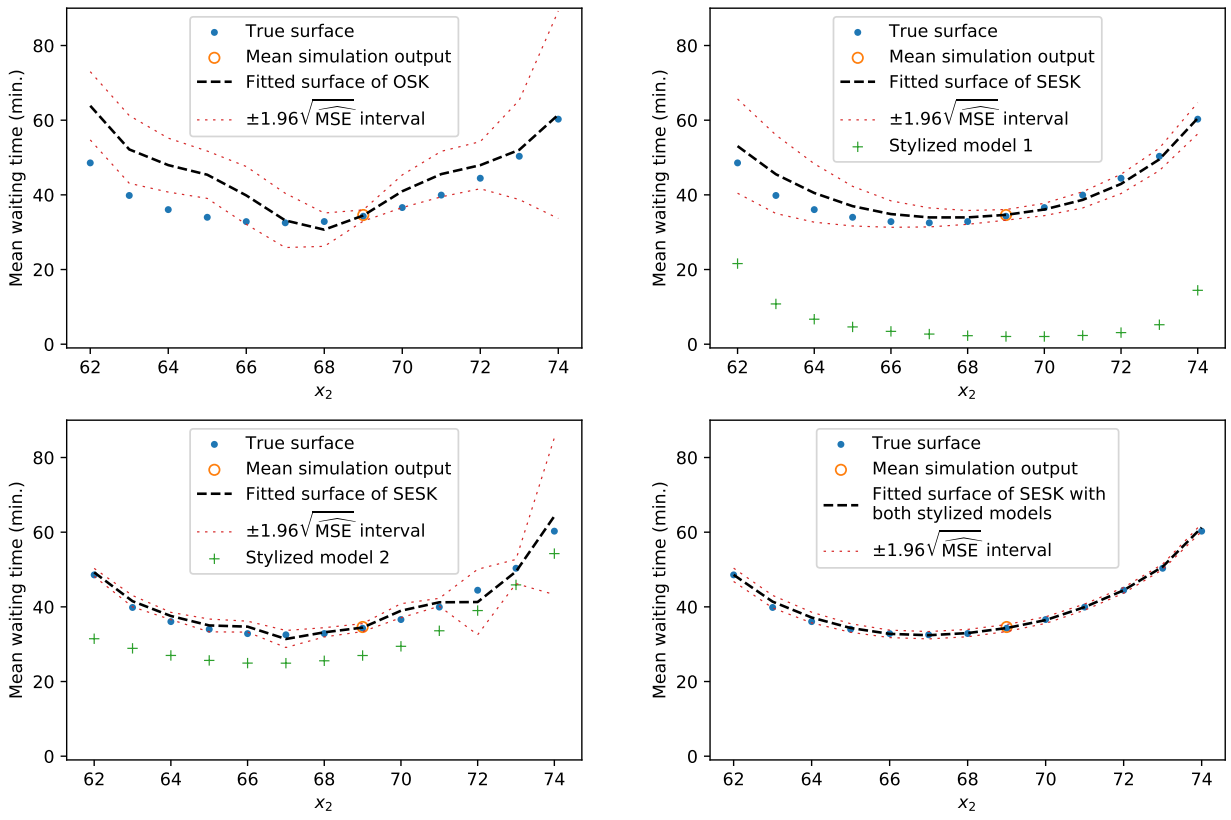

Figure 6: Predicted Trucks' Mean Waiting Time as a Function of x_2. Based on one typical design.

can further enhance the performance of the GPS algorithm.

Notice that stylized models can be integrated in the GPS algorithm by revising the algorithm so that it treats the residual surface $\mathcal{Y}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})^\intercal\boldsymbol{\beta}$ as a realization of a second-order stationary Gaussian process. Then, all the calculations involved in the GPS algorithm can be easily modified and we omit the details. We call the new algorithm stylized-model enhanced GPS (SEGPS).

To search for the optimal dock allocation, we use the fluid approximation as the stylized model, and initially set $\boldsymbol{\beta} = (0, 1)^\intercal$ and select 3 design points randomly, at each of which the simulation model is run to generate 5 independent observations. In each iteration of the SEGPS algorithm, we construct a Gaussian process based on all the previous simulation outputs. Then, we construct a sampling distribution and from it we sample the next 3 design points at each of which the simulation model is replicated for 5 times. For every 5 iterations (i.e., 15 design points), the value of $\boldsymbol{\beta}$ is updated through linear regression. The whole algorithm is stopped after 90 design points are visited, so the total number of observations is 450 upon termination. We repeat the experiment 40 times and compute the average estimated optimal value as a function of the number of observations. For comparison, the original GPS algorithm is also implemented in the same manner with the only difference that there is no stylized model $\boldsymbol{f}(\boldsymbol{x})$ (thus also no updating about $\boldsymbol{\beta}$). Figure 7 shows that the SEGPS algorithm converges significantly faster than the GPS algorithm.

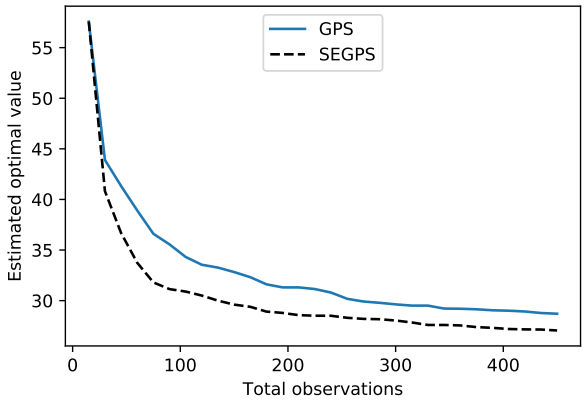

Figure 7: DOvS for Dock Allocation – GPS v.s. SEGPS.

7 Conclusions

We propose in this paper a simple, effective approach to improve the performance of stochastic kriging for queueing simulation. By incorporating stylized models that provide useful information about the shape of the unknown response surface, we demonstrate that the prediction accuracy can be improved substantially through two representative case studies. The stylized models need not to be highly accurate. Instead, the performance of SESK is fairly robust relative to the choice of the stylized models. This feature is particularly appealing to practitioners. Finally, we briefly illustrate that in addition to improving prediction accuracy, incorporation of stylized models can accelerate the convergence of the GPS algorithm for DOvS problems. We believe that the same idea should be able to extend to other algorithms for more general simulation optimization problems. We leave it to future investigation.

Acknowledgments

We thank Associate Editor and two anonymous referees for their constructive comments that improve the paper substantially. We gratefully acknowledge the support from the Hong Kong Research Grants Council under Project No. 9042362 and Project No. 16211417.

A Value Range of K^2

If $\boldsymbol{\Sigma} = \mathbf{0}$, the log-likelihood function (5) reduces to

$$\begin{aligned}\ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}) &= -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|\tau^2\boldsymbol{R}(\boldsymbol{\theta}) + \boldsymbol{\Sigma}| - \frac{1}{2}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta})^\intercal[\tau^2\boldsymbol{R}(\boldsymbol{\theta}) + \boldsymbol{\Sigma}]^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta}) \\ &= -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln(\tau^2) - \frac{1}{2}\ln|\boldsymbol{R}(\boldsymbol{\theta})| - \frac{1}{2\tau^2}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta})^\intercal\boldsymbol{R}(\boldsymbol{\theta})^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{\beta}).\end{aligned}$$

Further, if $\boldsymbol{\theta}$ is known, it is easy to see that the maximizer $(\widehat{\boldsymbol{\beta}}_S, \widehat{\tau}_S^2) = \arg\max_{\boldsymbol{\beta},\tau^2} \ell(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ satisfies

$$k\widehat{\tau}_S^2 = (\bar{\boldsymbol{Y}} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}_S)^\intercal \boldsymbol{R}(\boldsymbol{\theta})^{-1}(\bar{\boldsymbol{Y}} - \boldsymbol{F}\widehat{\boldsymbol{\beta}}_S).$$

It follows that

$$\ell(\widehat{\boldsymbol{\beta}}_S, \widehat{\tau}_S^2, \boldsymbol{\theta}) = -\frac{k}{2}(\ln(2\pi) + 1) - \frac{1}{2}\ln|\boldsymbol{R}(\boldsymbol{\theta})| - \frac{k}{2}\ln(\widehat{\tau}_S^2). \tag{18}$$

Moreover, notice that

$$\ell(\widehat{\boldsymbol{\beta}}_S, \widehat{\tau}_S^2, \boldsymbol{\theta}) = \max_{\beta_0,\beta_1,\tau^2} \ell((\beta_0, \beta_1), \tau^2, \boldsymbol{\theta}) \geq \max_{\beta_0,\tau^2} \ell((\beta_0, 0), \tau^2, \boldsymbol{\theta}) = \ell(\widehat{\boldsymbol{\beta}}_O, \widehat{\tau}_O^2, \boldsymbol{\theta}).$$

Then, by (18), we conclude that $0 \leq \widehat{\tau}_S^2 \leq \widehat{\tau}_O^2$, and thus $0 \leq K^2 \leq 1$.

B Decomposition of Finite-Capacity Queueing Networks

To facilitate the presentation of the stylized queueing model used in §6.1, we first introduce some notations. Let K be the number of stations in the network. For station i, let γ_i, λ_i, d_i, μ_i, and c_i denote its external arrival rate, internal arrival rate, departure rate, service rate, and number of servers, respectively, $i = 1, \ldots, K$. After service completion at station i, a customer is routed to station j with probability p_{ij} or leaves the network with probability $p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$.

The stylized queueing model for the finite-capacity queueing network with bufferless stations and BAS blocking mechanism in §6.1 is constructed as follows. Following the method in Korporaal et al. (2000), we decompose the network into K isolated independent stations, each of which is a queue of type $(M(\gamma)+M(\lambda))/M(\nu)/s/N$. The performance measure of the network is then approximated by aggregating that of each isolated station. Here, s is the number of servers, N is the buffer size, i.e. the maximum queue length, and $M(\nu)$ means the service time follows the exponential distribution with mean $1/\nu$. Moreover, $M(\gamma)$ and $M(\lambda)$ represent two independent Poisson arrival processes with arrival rates γ and λ, formed by the external and internal customer arrivals, respectively. We differentiate two types of customer loss. An external customer is considered a loss if he finds all s servers occupied upon his arrival. By contrast, an internal customer is considered a loss if he finds all s servers and all N waiting positions are occupied upon his arrival. Notice that the number of customers in this queueing model forms a finite-state birth-death process, so its steady-state performance measures can be calculated easily. We omit the details and present the results below.

Proposition 1. *Let $\{\pi_i : i = 0, 1, \ldots\}$ be the steady state distribution of the number of customers in an $(M(\gamma) + M(\lambda))/M(\nu)/s/N$ queue. Then,*

$$\pi_i = \begin{cases} \left(\frac{\gamma+\lambda}{\nu}\right)^i \frac{1}{i!}\pi_0, & \text{if } 0 \leq i \leq s, \\ \left(\frac{\gamma+\lambda}{\nu}\right)^s \frac{1}{s!}\left(\frac{\lambda}{s\nu}\right)^{i-s}\pi_0, & \text{if } s+1 \leq i \leq s+N \text{ and } N \geq 1, \end{cases}$$

where,

$$\pi_0 = \begin{cases} 1/\left[\sum_{i=0}^{s}\left(\frac{\gamma+\lambda}{\nu}\right)^i \frac{1}{i!} + \left(\frac{\gamma+\lambda}{\nu}\right)^s \frac{1}{s!}\sum_{i=1}^{N}\left(\frac{\lambda}{s\nu}\right)^i\right], & \text{if } N \geq 1, \\ 1/\left[\sum_{i=0}^{s}\left(\frac{\gamma+\lambda}{\nu}\right)^i \frac{1}{i!}\right], & \text{if } N = 0. \end{cases}$$

Let $L(\gamma,\lambda,\nu,s,N)$, $Q(\gamma,\lambda,\nu,s,N)$, $B_{\mathrm{E}}(\gamma,\lambda,\nu,s,N)$, and $B_{\mathrm{I}}(\gamma,\lambda,\nu,s,N)$ denote the mean number of customers in system, mean queue length, loss probability of external customers, and loss probability of internal customers in steady state, respectively. Then,

$$\begin{aligned} L(\gamma,\lambda,\nu,s,N) &= \sum_{i=1}^{s+N} i\pi_i, \\ Q(\gamma,\lambda,\nu,s,N) &= \begin{cases} \sum_{i=s+1}^{s+N}(i-s)\pi_i, & \text{if } N \geq 1, \\ 0, & \text{if } N = 0, \end{cases} \\ B_{\mathrm{E}}(\gamma,\lambda,\nu,s,N) &= \sum_{i=s}^{s+N} \pi_i, \\ B_{\mathrm{I}}(\gamma,\lambda,\nu,s,N) &= \pi_{s+N}. \end{aligned}$$

We now determine the parameters for each isolated $(M(\gamma)+M(\lambda))/M(\nu)/s/N$ queue. Following Korporaal et al. (2000), we give the following heuristic iterative algorithm. Linear interpolation (of performance measures) are applied to deal with non-integer s and N.

(i) Specify a small $\epsilon > 0$. Set $n = 0$ and specify an initial guess of the loss probability $b_j^{(0)}$ (e.g. 0). Let $\nu_j = \mu_j$, $s_j = c_j$, and $N_j = \sum_{i=1}^{K} c_i p_{ij}$, for $j = 1,\ldots,K$.

(ii) Update the parameters via the following equations

$$\begin{aligned} d_j &= \gamma_j(1-b_j) + \sum_{i=1}^{K} d_i p_{ij}, \quad j = 1,\ldots,K, \\ \lambda_j &= \sum_{i=1}^{K} d_i p_{ij}, \quad j = 1,\ldots,K, \\ s_j &= c_j - \sum_{i=1}^{K} \frac{d_j p_{ji}}{d_i} Q(\gamma_i,\lambda_i,\nu_i,s_i,N_i), \quad j = 1,\ldots,K, \\ b_j &= B_{\mathrm{E}}(\gamma_j,\lambda_j,\nu_j,s_j,N_j), \quad j = 1,\ldots,K. \end{aligned}$$

(iii) If $\max_{j=1,\ldots,K} |b_j - b_j^{(n)}| < \epsilon$, stop. Otherwise, let $b_j^{(n+1)} = \frac{b_j^{(n)}+b_j}{2}$ and $n \leftarrow n+1$; go to step (ii).

(iv) Compute ν_j for $j = 1, \ldots, K$ via

$$\frac{1}{\nu_j} = \frac{1}{\mu_j} + \sum_{i \in \{i | p_{ji} > 0\}} \frac{B_{\mathrm{I}}(\gamma_i, \lambda_i, \mu_i, s_i, N_i)\lambda_i}{d_j \mu_i s_i}.$$

After the parameters of each isolated $(M(\gamma) + M(\lambda))/M(\nu)/s/N$ queue are determined, the mean sojourn time S of the network can be computed via Little's law,

$$\sum_{j=1}^{K} L(\gamma_j, \lambda_j, \nu_j, s_j, N_j) = S \sum_{j=1}^{K} \gamma_j (1 - b_j).$$

C Fluid Approximation of Time-Varying Queues

We first use an $M_t/M/s$ queue as a simple approximation of the original $M_t/G/s$ queue in §6.2. Let $X(t)$ denote the number of customers in the system at time t. Then, $\bar{X}(t)$, the fluid approximation to $\mathbb{E}[X(t)]$, is available in closed form; see Gautam (2012, §8.4). Further, we propose the following heuristic approach for approximating the long-run mean waiting time of the time-varying queue. Suppose that the arrival rate is cyclic with period T. For example, in the dock allocation problem in §6.2, the arrival rates of cargoes to an air cargo terminal cycle with a 24-hour period. For a customer who arrives at time t, we approximate his expected waiting time in queue, $W_Q(t)$, by

$$W_Q(t) \approx \frac{1}{s\mu}[\bar{X}(t) - s + 1]^+, \tag{19}$$

where $[x]^+ = \max\{0, x\}$. The interpretation of this approximation is as follows. If there is at least one idle server upon his arrival, i.e. $X(t) \leq s - 1$, then the customer needs no waiting. Otherwise, he needs to wait for $X(t) - (s - 1)$ customers to depart the system. Since all the s servers are working, the departure rate is $s\mu$, so the waiting time is $[X(t) - (s - 1)]/(s\mu)$.

To approximate the long-run mean waiting time, we average the expected waiting times of all the customers that arrive during a cyclic period. Specifically, we apply a piecewise constant approximation for the arrival rate process. Suppose that the period $[0, T]$ is decomposed into L pieces, $\{[t_{\ell-1}, t_\ell) : \ell = 1, \ldots, L\}$, with $t_0 = 0$ and $t_L = T$. Suppose that piece ℓ has average arrival rate $\bar{\lambda}_\ell$. Then, the mean number of arrivals during $[t_{\ell-1}, t_\ell)$ is approximately $(t_\ell - t_{\ell-1})\bar{\lambda}_\ell$. We assume that the arrival times of these customers are evenly distributed on $[t_{\ell-1}, t_\ell)$ and let ξ_ℓ denote the collection of these arrival times, i.e.,

$$\xi_\ell = \left\{ t_{\ell-1} + i\frac{1}{\bar{\lambda}_\ell} \;\middle|\; i = 0, 1, \ldots, \lfloor (t_\ell - t_{\ell-1})\bar{\lambda}_\ell \rceil - 1 \right\},$$

where $\lfloor \cdot \rceil$ is the round function that returns the nearest integer. Then, we approximate the long-run

mean waiting time, W_Q, by

$$W_Q \approx \frac{1}{|\boldsymbol{\xi}|} \sum_{t \in \boldsymbol{\xi}} W_Q(t),$$

where $W_Q(t)$ is approximated by (19), $\boldsymbol{\xi} \coloneqq \bigcup_{\ell=0}^{L} \xi_\ell$ and $|\boldsymbol{\xi}|$ denotes its cardinality.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control 19*(6), 716–723.

Andradóttir, S. (1996). A global search method for discrete stochastic optimization. *SIAM J. Optim. 6*(2), 513–530.

Andradóttir, S. (2006). An overview of simulation optimization via random search. In *Handbooks in Operations Research and Management Science*, Volume 13, pp. 617–631. Elsevier.

Ankenman, B., B. L. Nelson, and J. Staum (2010). Stochastic kriging for simulation metamodeling. *Oper. Res. 58*(2), 371–382.

Asmussen, S. (1992). Queueing simulation in heavy traffic. *Math. Oper. Res. 17*(1), 84–111.

Asmussen, S. (2008). *Applied Probability and Queues*. Springer.

Balsamo, S., V. de Nitto Personé, and R. Onvural (2013). *Analysis of Queueing Networks with Blocking*. Springer.

Barton, R. R. (1998). Simulation metamodels. In *Proc. 1998 Winter Simulation Conf.*, pp. 167–174.

Billingsley, P. (1999). *Convergence of Probability Measures* (2 ed.). Wiley-Interscience.

Chao, X. (2011). Networks with customers, signals, and product form solutions. In R. J. Boucherie and N. M. van Dijk (Eds.), *Queueing Networks: A Fundamental Approach*, pp. 217–267. Springer.

Chen, X., B. Ankenman, and B. L. Nelson (2013). Enhancing stochastic kriging metamodels with gradient estimators. *Oper. Res. 61*(2), 512–528.

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

Forrester, A. I. J., A. Sóbester, and A. J. Keane (2007). Multi-fidelity optimization via surrogate modelling. In *Proc. R. Soc. A*, Volume 463, pp. 3251–3269.

Forrester, A. I. J., A. Sóbester, and A. J. Keane (2008). *Engineering Design via Surrogate Modelling: A Practical Guide*. John Wiley & Sons, Inc.

Gautam, N. (2012). *Analysis of Queues: Methods and Applications*. CRC Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *J. Glob. Optim. 34*(3), 441–466.

Jackson, J. R. (1963). Jobshop-like queueing systems. *Manag. Sci. 10*(1), 131–142.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley & Sons, Inc.

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

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *J. R. Statist. Soc. B 63*(3), 425–464.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *Eur. J. Oper. Res. 192*(3), 707–716.

Korporaal, R., A. Ridder, P. Kloprogge, and R. Dekker (2000). An analytic model for capacity planning of prisons in the Netherlands. *J. Oper. Res. Soc. 51*(11), 1228–1237.

Lin, Z., A. Matta, N. Li, and J. G. Shanthikumar (2016). Extended kernel regression: A multi-resolution method to combine simulation experiments with analytical methods. In *Proc. 2016 Winter Simulation Conf.*, pp. 590–601.

Mardia, K. V. and R. J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika 71*(1), 135–146.

Matheron, G. (1963). Principles of geostatistics. *Econ. Geol. 58*(8), 1246–1266.

Mitchell, T., M. Morris, and D. Ylvisaker (1994). Asymptotically optimum experimental designs for prediction of deterministic functions given derivative information. *J. Stat. Plann. Infer. 41*(3), 377–389.

Morris, M. D. and T. J. Mitchell (1995). Exploratory designs for computational experiments. *J. Stat. Plann. Infer. 43*(3), 381–402.

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.

Osorio, C. and M. Bierlaire (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *Eur. J. Oper. Res. 196*(3), 996–1007.

Qu, H. and M. C. Fu (2014). Gradient extrapolated stochastic kriging. *ACM Trans. Model. Comput. Simul. 24*(4), 23:1–23:25.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat. 6*(2), 461–464.

Scott, W., P. Frazier, and W. Powell (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM J. Optim. 21*(3), 996–1026.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. 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. *Oper. Res. 62*(6), 1416–1438.

Whitt, W. (1989). Planning queueing simulations. *Manag. Sci. 35*(11), 1341–1366.

Whitt, W. (1991). The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Manag. Sci. 37*(3), 307–314.

Xie, W., B. L. Nelson, and R. R. Barton (2014). A Bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res. 62*(6), 1439–1452.

Xie, W., B. L. Nelson, and J. Staum (2010). The influence of correlation functions on stochastic kriging metamodels. In *Proc. 2010 Winter Simulation Conf.*, pp. 1067–1078.

Yang, F., B. Ankenman, and B. L. Nelson (2007). Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Res. Logist. 54*(1), 78–93.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc. 99*(465), 250–251.

Zhang, H. and D. L. Zimmerman (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika 92*(4), 921–936.