Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Optimal Simulator Selection

Ying Hung, Li-Hsiang Lin & C. F. Jeff Wu

To cite this article: Ying Hung, Li-Hsiang Lin & C. F. Jeff Wu (2023) Optimal Simulator Selection, Journal of the American Statistical Association, 118:542, 1264-1271, DOI: 10.1080/01621459.2021.1987920

To link to this article: https://doi.org/10.1080/01621459.2021.1987920

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 542, 1264–1271: Theory and Methods
https://doi.org/10.1080/01621459.2021.1987920

Optimal Simulator Selection

Ying Hung[a], Li-Hsiang Lin[b], and C. F. Jeff Wu[c]

[a]Department of Statistics, Rutgers University, Newark, NJ; [b]Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA; [c]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

ABSTRACT

Computer simulators are widely used for the study of complex systems. In many applications, there are multiple simulators available with different scientific interpretations of the underlying mechanism, and the goal is to identify an optimal simulator based on the observed physical experiments. To achieve the goal, we propose a selection criterion based on leave-one-out cross-validation. This criterion consists of a goodness-of-fit measure and a generalized degrees of freedom penalizing the simulator sensitivity to perturbations in the physical observations. Asymptotic properties of the selected optimal simulator are discussed. It is shown that the proposed procedure includes a conventional calibration method as a special case. The finite sample performance of the proposed procedure is demonstrated through numerical examples. In the application of cell biology, an optimal simulator is selected, which can shed light on the T cell recognition mechanism in the human immune system. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2020
Accepted September 2021

KEYWORDS
Calibration; Computer experiment; Cross-validation; Emulation; Generalized degrees of freedom

1. Introduction

There are generally two types of experiments for the studies of complex systems: physical and computer experiments. Physical experiments refer to actual experiments performed in a laboratory or observed in the field. They are often expensive and/or infeasible to conduct. Therefore, computer experiments, which refer to simulations using complex mathematical models and numerical tools, are commonly served as alternatives.

Based on the observed physical experiments, there is a growing demand in many applications for the identification of an optimal computer simulator among multiple ones with different scientific interpretations of the underlying mechanisms. For example, among different queuing models which represent different types of patient flow, it is important to identify the best simulator for a particular medical service in a hospital (Lakshmi and Iyer 2013). Geologists want to know that, among different global weather models governed by different fluid dynamics and thermodynamics equations, which one can be best used for predicting the weather of a local region (Richardson 2007). Biologists need to select some differential equations to represent the growth (or decline) of a biological population (Brauer and Castillo-Chavez 2012). However, most of the existing developments in the computer experiment literature are based on the assumption that there is only one simulator available (Fang, Li, and Sudjianto 2006; Santner et al. 2018). To the best of our knowledge, there is no systematic procedure developed for the selection of an optimal simulator.

The goal in optimal simulator selection is different from the variable selection problems in computer experiments (Bastos and O'Hagan 2009; Overstall and Woods 2016), where the focus is to identify significant variables by using one computer simulator. It is also different from studies of multi-fidelity simulations where multiple simulations are developed based on the same physical law but with different approximation accuracy, and the objective is to incorporate information efficiently from all the computer simulators (Kennedy and O'Hagan 2000; Tuo, Wu, and Yu 2014).

To identify an optimal simulator, we propose a new criterion based on leave-one-out cross-validation (LOOCV). LOOCV is commonly used for model selection, but how to implement this idea to select simulators is not trivial because of two reasons. First, unlike typical model selection problems, the current setting involves two types of data, one from physical experiments and another from computer simulations. Second, in addition to a set of regular parameters shared across all the simulations, each simulator is associated with a unique set of calibration parameters. These two types of parameters play different roles, and it is crucial to distinguish their impacts in the optimal simulator selection. The proposed LOOCV criterion addresses these issues by incorporating a goodness-of-fit measure for each simulator and a generalized degrees of freedom penalizing the sensitivity of the simulator due to calibration. Different from typical AIC and BIC types of methods (Burnham and Anderson 2011; Wood, Pya and Säfken 2016), where the penalty is defined directly by the total number of parameters, the proposed criterion takes into account the unique feature of calibration and offers a data-driven penalty for the simulator sensitivity.

This article is organized as follows. In Section 2, we propose a criterion for selecting the optimal simulator. In Section 3, the proposed criterion is shown to be decomposed into a measure of goodness of fit for physical experiments and a generalized degrees of freedom which properly penalizes the sensitivity of the simulator due to the calibration parameters. It is also

CONTACT C. F. Jeff Wu jeffwu@isye.gatech.edu School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

shown that the proposed criterion includes the conventional L_2-norm calibration criterion (Tuo and Wu 2015) as a special case when there is only one simulator available. The asymptotic properties of the selected optimal simulator are also discussed in Section 3. The simulation studies are provided in Section 4, and an application of the proposed method in selecting the optimal antigen recognition mechanism for T-cell signaling is provided in Section 5.

2. Cross-Validation for Optimal Simulator Selection

Assume that there are n observations available from physical experiments denoted by $\mathbf{D} \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i$ is a p-dimensional input. For notational simplicity, we first assume the outputs y_i's are continuous, and

$$y(\mathbf{x}_i) = \xi(\mathbf{x}_i) + \epsilon_i, \tag{1}$$

where $y(\mathbf{x}_i) = y_i, \xi(\mathbf{x}_i)$ is known as the *true process* in computer experiment literature, and ϵ_i are identically distributed random variables with zero mean and finite variance (Tuo and Wu 2015). The true process ξ can be estimated by nonparametric regression methods, such as kernel ridge regression (Shawe-Taylor and Cristianini 2004) and Gaussian process (Fang, Li, and Sudjianto 2006; Santner et al. 2018), and the estimator is denoted by $\hat{\xi}(\cdot)$. Apart from physical experiments, there are K candidate computer simulators which refer to K different mathematical models representing different underlying mechanisms. These simulators are often computationally intensive to perform, especially for the studies of complex systems; therefore, they are approximated by surrogate models for further analysis and inference. These surrogate models are known as *emulators* denoted by $f_k(\mathbf{x}, \boldsymbol{\theta}_k, \boldsymbol{\beta}_k)$, where $k = 1, \ldots, K$, $\boldsymbol{\theta}_k$ is a set of unknown parameters called calibration parameters (Santner et al. 2018), and $\boldsymbol{\beta}_k$ is the rest of the unknown parameters for constructing the kth emulator. The parameters, $\boldsymbol{\theta}_k$ and $\boldsymbol{\beta}_k$, can be different over k. Various surrogate models, such as Gaussian process models or spline-based models (Wahba 1990), are applicable here to construct emulators, and the parameters $\boldsymbol{\beta}_k$ can be estimated accordingly. Given the computer experiment outputs $\mathbf{D}_k^s$ collected from the kth simulator, we assume that one emulator, $f_k(\mathbf{x}; \boldsymbol{\theta}_k, \hat{\boldsymbol{\beta}}_k)$, is constructed and serves as the surrogate for the simulator to perform prediction, inference, and uncertainty quantification. By incorporating the information from the physical experiments $\mathbf{D}$, the calibration parameter $\hat{\boldsymbol{\theta}}_k(\mathbf{D})$ can be estimated by calibration methods, including the L_2 calibration which minimizes the discrepancy between the simulator and the true process $\xi(\cdot)$ (Tuo and Wu 2015) and the least-square approach which minimizes the least-square distance between the simulator and the true process.

Given the outputs from the K simulators and the observations from physical experiments, the goal is to identify an *optimal emulator* $f_0(\mathbf{x}; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$ satisfying

$$f_0(\mathbf{x}; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \underset{f_1, f_2, \ldots, f_K}{\arg\min} \{||\xi(\mathbf{x}) - f_k(\mathbf{x}; \boldsymbol{\theta}_k, \boldsymbol{\beta}_k)||_{L_2}\}, \tag{2}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\beta}_k$ are the true parameter settings associated with the surrogate model f_k and a prespecified calibration procedure, and $||\cdot||_{L_2}$ is the L_2 norm. We call the corresponding simulator of the optimal emulator the *optimal simulator*. To estimate (2), we propose a leave-one-out cross-validation (LOOCV) method as follows. Define a LOOCV score by

$$\widehat{\mathrm{Err}}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathrm{Err}}_{k,(i)}, \tag{3}$$

where $\mathbf{D}_{(-i)} = \mathbf{D} \backslash \{(\mathbf{x}_i, y_i)\}$, $\widehat{\mathrm{Err}}_{k,(i)} \equiv Q\left(\hat{\xi}(\mathbf{x}_i), f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}), \hat{\boldsymbol{\beta}}_k(\mathbf{D}_{(-i)}))\right)$, $Q(\cdot, \cdot)$ is a prespecified loss function, $\hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)})$ is the estimated calibration parameters by using dataset $\mathbf{D}_{(-i)}$, and $\hat{\xi}(\cdot)$ is the estimated true process. To guarantee the theoretical properties in Section 3, $\hat{\boldsymbol{\theta}}_k(\cdot)$ is required to be $\sqrt{n}$-consistent estimators which are obtainable by several calibration methods including those discussed in Tuo and Wu (2015), Wong, Storlie, and Lee (2017), and Sung et al. (2020a) and the necessary condition for $\hat{\xi}(\cdot)$ can be achieved by commonly used nonparametric methods, such as kernel ridge regression methods and Gaussian processes (Stone 1982; Tsybakov 2008). In this article, we consider three types of loss functions for $Q(\cdot, \cdot)$ including the squared loss, the zero-one loss, and the deviance loss (Efron 1986; Gneiting and Raftery 2007). Because $\hat{\boldsymbol{\beta}}_k$ is estimated from $\mathbf{D}_k^s$, we have $\hat{\boldsymbol{\beta}}_k(\mathbf{D}_{(-1)}) = \cdots = \hat{\boldsymbol{\beta}}_k(\mathbf{D}_{(-n)})$. Therefore, with a slight abuse of notation, $\hat{\boldsymbol{\beta}}_k(\mathbf{D}_{(-i)})$ is omitted in the cross-validation iterations and $f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}), \hat{\boldsymbol{\beta}}_k(\mathbf{D}_{(-i)}))$ is replaced by $f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}))$ for notation simplicity.

Based on Equation (3), we obtain the estimated optimal emulator

$$f_T(\mathbf{x}; \hat{\boldsymbol{\theta}}_T(\mathbf{D})) \text{ with } T \equiv \underset{k=1,\ldots,K}{\arg\min} \widehat{\mathrm{Err}}_k, \tag{4}$$

where $\hat{\boldsymbol{\theta}}_T(\mathbf{D})$ is the estimated calibration parameter from $\mathbf{D}$. The procedure is summarized in Algorithm 1. This procedure can also be generalized to non-Gaussian outputs. Take the binary output as an example, the same procedure follows by replacing the true process by $\xi(\mathbf{x}) = P(y(\mathbf{x}) = 1)$. A demonstration of Algorithm 1 in the application to binary output is given in Section 5.

The proposed procedure assumes that each simulator is represented by one emulator. In practice, there are numbers of surrogate models that can be used to construct emulators; therefore, it is crucial to carefully select one of them to represent the simulator. To do so, one approach is to modify the proposed LOOCV procedure by replacing f_k in Algorithm 1 with different surrogate models and find the one that minimizes the $\widehat{\mathrm{Err}}_k$. Furthermore, an accurate estimation of the true process, $\hat{\xi}(\cdot)$, is crucial as shown in the next section. To further enhance the asymptotic performance, $\hat{\xi}(\cdot)$ can be incorporated into the proposed LOOCV procedure by estimating $\xi(\cdot)$, denoted by $\hat{\xi}(\mathbf{x}_i; \mathbf{D}_{(-i)})$, after line 5 of Algorithm 1 instead of line 2.

3. Theoretical Properties

In the following lemma, it is shown that the LOOCV score in Equation (3) can be decomposed into the goodness-of-fit

Algorithm 1 The algorithm for simulator selection

procedure LOOCV($\mathbf{D} \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^n$), $\{\mathbf{D}_k^s : k = 1, \ldots, K\}$)
  Estimate the true process $\hat{\xi}$
  for each k in $1, 2, \ldots, K$ **do**
    Construct emulator $f_k(\mathbf{x}; \boldsymbol{\theta}_k, \hat{\boldsymbol{\beta}}_k)$ using $\mathbf{D}_k^s$.
    for each i in $1, 2, \ldots, n$ **do**
      Obtain the estimated calibration parameter $\hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)})$.
      Calculate $\widehat{\text{Err}}_{k,(i)} = Q(\hat{\xi}(\mathbf{x}_i), f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)})))$
    end for
    Obtain $T \equiv \arg\min_{k=1,\ldots,K} \widehat{\text{Err}}_k$, where $\widehat{\text{Err}}_k = \frac{1}{n}\sum_{i=1}^n \widehat{\text{Err}}_{k,(i)}$.
  end forreturn The selected optimal simulator f_T.
end procedure

of the emulator and a quantity penalizing the flexibility of the emulator. Therefore, minimizing (3) implies a minimization of not only the discrepancy between physical and computer experiments but also the sensitivity of the emulator. The detailed proofs can be found in the supplemental material Section 2.

The decomposition of Equation (3) requires an expression of the loss function which is introduced by Efron (1986). That is,

$$Q(\hat{\xi}(\mathbf{x}_i), f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}))) = q(f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}))) + \dot{q}(f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D})))\{\hat{\xi}(\mathbf{x}_i) - f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}))\}, \tag{5}$$

where $q(\cdot)$ is a concave function and $\dot{q}(\cdot)$ is its first-order derivative. For example, as shown in the supplementary material Section 6, a squared loss function $Q(\cdot, \cdot)$ can be expressed by Equation (5) with $q(f(\mathbf{x})) = f(\mathbf{x})(y(\mathbf{x}) - f(\mathbf{x}))$.

Lemma 1.

$$E[\widehat{\text{Err}}_k] = \bar{\text{err}}_k + \frac{1}{n} GD_k, \tag{6}$$

where

$$\bar{\text{err}}_k = \frac{1}{n}\sum_{i=1}^n Q(y_i, f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}))) \tag{7}$$

and

$$\text{GD}_k = \sum_{i=1}^n \text{cov}\left(-\dot{q}\left[f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}))\right], y_i - \hat{\xi}(\mathbf{x}_i)\right). \tag{8}$$

The quantity in Equation (7) determines how well the emulator fits the physical observations and the quantity GD_k in (8) is the *generalized degrees of freedom* for the kth emulator which is an analogy to "optimism" in Efron (1986) and the generalized degrees of freedom for linear model in Ye (1998). The quantity GD_k can be estimated by $\widehat{\text{GD}}_k = n(\widehat{\text{Err}}_k - \bar{\text{err}}_k)$ and can be interpreted as the sum of sensitivity of the kth estimated emulator to perturbations in the corresponding physical observation $y_i - \hat{\xi}(\mathbf{x}_i)$. If the emulator is highly flexible/sensitive, then the values in f_k tend to have a higher correlation with $y_i - \hat{\xi}(\mathbf{x}_i)$, which leads to a larger penalty. It also appears that the sensitivity is mainly associated with calibration because $\text{GD}_k = \sum_{i=1}^n \text{cov}\left(-\dot{q}\left[f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k(\mathbf{D}_{(-i)}))\right], y_i - \hat{\xi}(\mathbf{x}_i)\right) = \sum_{i=1}^n -\dot{q}\left[f_k(\mathbf{x}_i)\right] E\left\{(y_i - \hat{\xi}(\mathbf{x}_i))\right\} = 0$ if there is no calibration parameters involved in the kth emulator. As compared to a naive application, where LOOCV is evaluated based on the physical observations y_i directly instead of $\hat{\xi}$ in (3), the value of GD_k is always zero which indicates no penalty for the sensitivity. This result demonstrates the novelty of the proposed LOOCV where a data-driven penalty function is incorporated and the penalty function automatically distinguish the impacts of calibration from regular parameter estimation. It is shown in the following special case that the generalized degrees of freedom is equivalent to the number of calibration parameters. The detailed proof is given in supplemental material Section 3.

Proposition 2. Suppose $q(\cdot) = -x^2/2$, $f_k(\mathbf{x}_i; \boldsymbol{\theta}_k) = \mathbf{x}_i^T \boldsymbol{\theta}_k$ and $y_i = \mu_i + \epsilon_i$, where ϵ_i are i.i.d standard normal for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. If $\boldsymbol{\theta}_k$ is estimated by least-square method, the generalized degree of freedom GD_k in (8) is equal to the dimension of $\boldsymbol{\theta}_k$.

The proposed selection procedure can also be applied to the conventional calibration problem with $K = 1$. The following result shows that the estimated calibration parameters and the resulting discrepancy based on the proposed leave-one-out procedure asymptotically converge in probability to those obtained by the conventional L_2 calibration (Tuo and Wu 2015; Sung et al. 2020a). The proof is given in supplemental Material Section 4.

Theorem 3. Suppose $Q(\cdot, \cdot)$ is the squared loss, and $\mathbf{x}$ follows a uniform distribution on $[0, 1]^p$ and $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\beta}}_k$ are $\sqrt{n}$-consistent estimators of $\boldsymbol{\theta}_k$ and $\boldsymbol{\beta}_k$. Suppose $\xi(\cdot)$ is a d-times differentiable function, and $||\hat{\xi}(\cdot) - \xi(\cdot)||_{L_2}$ is $o_p(n^{-d/(2d+p)})$. Under Assumption (A1) given in supplemental material Section 1, we have

(i)

$$\widehat{\text{Err}}_k = \frac{1}{n}\sum_{i=1}^n \widehat{\text{Err}}_{k,(i)} = ||\xi(\mathbf{x}) - f_k(\mathbf{x}; \boldsymbol{\theta}_k, \boldsymbol{\beta}_k))||_{L_2} + o_p(n^{-d/(2d+p)}). \tag{9}$$

(ii) When $K = 1$, $\widehat{\text{Err}}_1$ converges in probability to the minimum L_2 discrepancy defined by Tuo and Wu (2015) with convergent rate $o_p(n^{-d/(2d+p)})$.

For the estimated optimal emulator, its estimated prediction error and the sensitivity are denoted by $\widehat{\text{Err}}_T$ and $\widehat{\text{GD}}_T = n(\widehat{\text{Err}}_T - \bar{\text{err}}_T)$. For the optimal emulator $f_0(\mathbf{x}, \boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$ defined in (2), we denote its prediction error by $\text{Err}_0 = \frac{1}{n}\sum_{i=1}^n Q(\hat{\xi}(\mathbf{x}_i), f_0(\mathbf{x}_i; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0))$ and the corresponding sensitivity by $\text{GD}_0 = \sum_{i=1}^n \text{cov}\left(-\dot{q}\left[f_0(\mathbf{x}_i; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\right], y_i - \hat{\xi}(\mathbf{x}_i)\right)$. In the following theorem, it is shown that the estimated prediction error $\widehat{\text{Err}}_T$ and the estimated $\widehat{\text{GD}}_T$ are asymptotically equivalent to those calculated for the optimal emulator $f_0(\mathbf{x}, \boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$. The proof is given in supplemental material Section 5.

Theorem 4. Under the assumptions in Theorem 3.3 with Assumptions (A1)–(A3) in supplemental material Section 1. We have

(i) $\widehat{\text{Err}}_T = \text{Err}_0 + o_p(n^{-d/(2d+p)})$, and
(ii) $\widehat{\text{GD}}_T/n = \text{GD}_0/n + o_p(n^{-d/(2d+p)})$.

4. Numerical Studies

In this section, the true process $\xi(\mathbf{x})$ is estimated by the kernel ridge regression, that is, to minimize the following loss function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \xi(\mathbf{x}_i))^2 + \lambda||\xi||^2_{\mathcal{N}_\Phi}, \tag{10}$$

where $\lambda > 0$ is a penalized parameter, $||\cdot||_{\mathcal{N}_\Psi}$ is the norm of the reproducing kernel Hilbert space $\mathcal{N}_\Psi$ generated by a kernel function $\Psi(\cdot)$. We consider a Gaussian kernel $\Psi(h) = \exp(-h^2/(2\tau^2))$, where τ is the length scale parameter. The penalized parameter λ in Equation (10) and the length scale parameter τ are chosen by 10-fold cross-validation. The emulators are then constructed by the Gaussian process (GP) models

$$f_k(\mathbf{x};\boldsymbol{\theta}_k,\boldsymbol{\beta}_k) \sim \text{GP}(\mu_k, \sigma_k^2\Phi((\mathbf{x}',\boldsymbol{\theta}_k'),(\mathbf{x},\boldsymbol{\theta}_k))), \tag{11}$$

where $\Phi((\mathbf{x}',\boldsymbol{\theta}_k'),(\mathbf{x},\boldsymbol{\theta}_k))$ a Matèrn kernel with roughness coefficient 2.5 for kth emulator, i.e., $\Phi((\mathbf{x}',\boldsymbol{\theta}_k'),(\mathbf{x},\boldsymbol{\theta}_k)) = (1/[\Gamma(\nu)2^{\nu-1}])(\sqrt{2\nu}\{||\mathbf{x}' - \mathbf{x}||_2^2 + ||\boldsymbol{\theta}_k' - \boldsymbol{\theta}_k||_2^2\}/\rho_k)^\nu \mathcal{K}_\nu(\sqrt{2\nu}\{||\mathbf{x}' - \mathbf{x}||_2^2 + ||\boldsymbol{\theta}_k' - \boldsymbol{\theta}_k||_2^2\}/\rho_k)$ with $\nu = 2.5$, where $\Gamma(\cdot)$ is the gamma function, $\mathcal{K}_\nu(\cdot)$ is the Bessel function, and ρ_k is the range parameter. The model parameter $\boldsymbol{\beta}_k = (\mu_k, \sigma_k^2, \rho_k)$ in (11) includes the unknown mean, variance, and the range parameter for kth emulator, and is estimated by empirical maximum likelihood method (Santner et al. 2018, sec. 3.3). The calibration parameters are estimated by the L_2-calibration method (Tuo and Wu 2015).

4.1. Example 1: The Branin Function

Two simulators are constructed by the Branin function with two different sets of calibration parameters:

$$\begin{aligned} m_1(x_1,x_2;\boldsymbol{\theta}_1) &= \left(x_2 - b(\frac{x_1}{\pi})^2 + 5.5\frac{x_1}{\pi} - r\right)^2 \\ &\quad +1\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 1, \\ m_2(x_1,x_2;\boldsymbol{\theta}_2) &= \left(x_2 - b(\frac{x_1}{\pi})^2 + c\frac{x_1}{\pi} - 6\right)^2 \\ &\quad +1\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 1, \end{aligned} \tag{12}$$

where simulator m_1 contains the calibration parameters $\boldsymbol{\theta}_1 = (b, r)$, simulator m_2 contains the calibration parameters $\boldsymbol{\theta}_2 = (b, c)$, $b \in [0,2]$, $r \in [5,7]$, and $c \in [4,6]$. For both simulators, computer experiments are conducted by using a 60-run maximum projection design (Joseph, Gul, and Ba 2015). Simulator m_2 is used as the true process to generate physical experiments by $y(x_1,x_2) = m_2(x_1,x_2;\boldsymbol{\theta}_2) + \epsilon$, where the inputs x_1 and x_2 are 30 Sobol′ points, the calibration parameters are set to be $\boldsymbol{\theta}_2 = (1.275, 5)$, and $\epsilon \sim N(0,4)$.

The true process is estimated by minimizing the loss function (10) with the length scale parameter 2.635, selected by a 10-fold cross-validation. Based on Equation (3) and Lemma 1, the leave-one-out cross-validation scores for the two simulators are reported in Table 1 with the estimated generalized degrees of freedom. By using the proposed criterion, the selected optimal simulator is $T = 2$, which agrees with the numerical settings. Furthermore, the estimated generalized degrees of freedom for the two simulators are similar, which implies a similar sensitivity due to the calibration parameters for the two simulators. This observation also agrees with the numerical settings in which equal number of calibration parameters are associated with the simulators.

Table 1. The leave-one-out cross-validation scores and the estimated generalized degrees of freedom for the two simulators in Example 1.

k	$\widehat{\text{Err}}_k$	$\widehat{\text{GD}}_k$
1	7.878	3.016
2	6.469	3.138

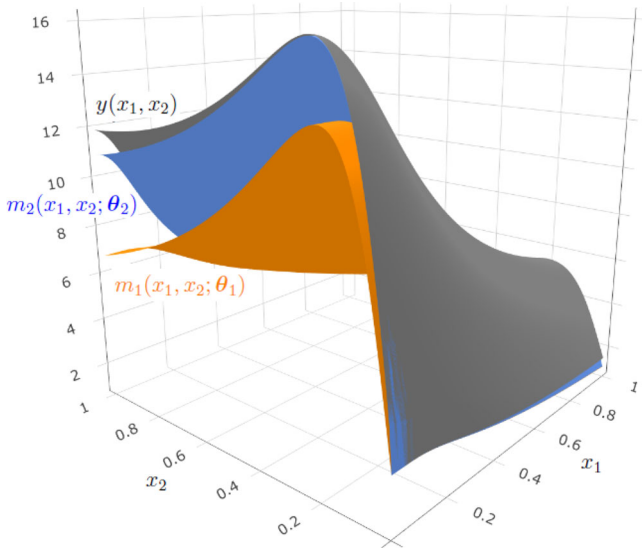

Figure 1. The response surfaces for the three functions in Example 2.

4.2. Example 2: Multi-Fidelity Simulators

The proposed procedure is demonstrated by using two simulators introduced by Goh et al. (2013) for the study of multi-fidelity simulations. Define the low-fidelity and high-fidelity simulators, m_1 and m_2, by

$$\begin{aligned} m_1(x_1,x_2;\boldsymbol{\theta}_1) &= \left(1 - \exp\left(\frac{1}{-2x_2}\right)\right) \\ &\quad \times \frac{1000t_s x_1^3 + 1900x_1^2 + 2092x_1 + 60}{1000t_\ell x_1^3 + 500x_1^2 + 4x_1 + 20}, \\ m_2(x_1,x_2;\boldsymbol{\theta}_2) &= m_1(x_1,x_2;\boldsymbol{\theta}_1) \\ &\quad +5\exp(-t_s)\frac{x_1^{t_h}}{100x_2^{2+t_h} + 1}, \end{aligned} \tag{13}$$

where the calibration parameters $\boldsymbol{\theta}_1 = (t_s, t_\ell) \in [0,1]^2$ and $\boldsymbol{\theta}_2 = (t_s, t_h) \in [0,1]^2$. When the high-fidelity simulator (13) is used, the calibration parameter t_ℓ in $\boldsymbol{\theta}_1$ is set to be 0.1. The physical experiments are generated by

$$y(x_1,x_2) = m_2(x_1,x_2;\boldsymbol{\theta}_2) + \frac{10x_1^2 + 4x_2^2}{50x_1x_2 + 10} + \epsilon$$

with the calibration parameters set to be $\boldsymbol{\theta}_2 = (0.2, 0.3)$ and $\epsilon \sim N(0, 0.25)$. These functions are shown in Figure 1. The goal is to

Table 2. The leave-one-out cross-validation scores and the estimated generalized degrees of freedom for the two simulators in Example 2.

k	$\widehat{\text{Err}}_k$	$\widehat{\text{GD}}_k$
1	1.429	4.808
2	0.283	5.200

identify the optimal simulator based on the observed physical data. This is different from the conventional goal in the study of multi-fidelity simulations.

A 40-run maximum projection design is used for the two simulators, and the physical experiments are performed based on a 30-run Sobol' points. The true process is estimated by minimizing (10) with the length scale parameter 0.584, selected by 10-fold cross-validation. For the two simulators, the estimated LOOCV scores and the estimated generalized degrees of freedom are reported in Table 2. Because $\widehat{\text{Err}}_1 > \widehat{\text{Err}}_2$, the high-fidelity simulator m_2 is chosen as the optimal simulator according to Equation (4). Based on Table 2, the high-fidelity simulator has a larger $\widehat{\text{GD}}_k$ as compared to the low-fidelity simulator which indicates a slightly higher flexibility and therefore a larger penalty for the sensitivity.

4.3. Example 3: The Study of Simulator Sensitivity

To demonstrate the performance of the generalized degrees of freedom with respect to the different sensitivity in simulators, we consider two simulators with different numbers of calibration parameters: $m_1(x;\boldsymbol{\theta}_1) = \delta_1 x + \delta_2 x^2 + \delta_3 x^3$ and $m_2(x;\theta_2) = \theta_2 x$, where $\boldsymbol{\theta}_1 = (\delta_1, \delta_2, \delta_3)$, and θ_2 are the calibration parameters. Physical experiments are generated from $y(x) = x + 2x^2 + 3x^3 + 0.1\sin(20x) + \epsilon$, where $\epsilon \sim N(0, 0.25)$. These functions are illustrated in Figure 2(a). A 100-run maximum projection design is used to generate computer experiments based on the two simulators, and a 61-run maximin design is implemented for physical experiments.

Based on 100 replicates, the average of $\widehat{\text{GD}}_1$ is 3.143 with standard deviation 0.342, and the average of $\widehat{\text{GD}}_2$ is 1.294 with standard deviation 0.154. These results are summarized in the boxplots in Figure 2(b). The estimated generalized degrees of freedom for the first simulator are around three times more than that of the second simulator, which reflects the sensitivity associated with the first simulator due to a larger number of calibration parameters and a higher-order polynomial.

5. Optimal Simulator for T-cell Signaling

It has long been known that the adaptive immune system defends the organism against diseases by recognition of pathogens by the T cell. T- cell receptor (TCR) is the primary molecule on the T cell in detecting foreign antigens which are present in major histocompatibility complex (pMHC) molecule expressed by infected cells. However, much is still unknown regarding the underlying antigen recognition mechanism.

To understand the recognition mechanism through the TCR-pMHC interactions, biologists (Rittase 2018) have developed micropipette adhesion frequency assays which are physical experiments performed in a laboratory. Although micropipette assays allow accurate measurements, they are time-consuming and often involve complicated experimental manipulation. Furthermore, some variables of interest cannot be studied in the lab due to technical complexity in carrying out the experiments. A more cost-effective approach is to illuminate the unknown recognition mechanism through computer simulations. Based on the idea of the kinetic proofreading model, two simulators are developed under two different recognition mechanisms: one is the conformation-change mechanism (denoted by CC in Figure 3(a)), and the other is the receptor-pulling mechanism (denoted by RP in Figure 3(b)). The two mechanisms associate with two different ways of TCR-pMHC interactions, either the molecules have conformational change due to the binding

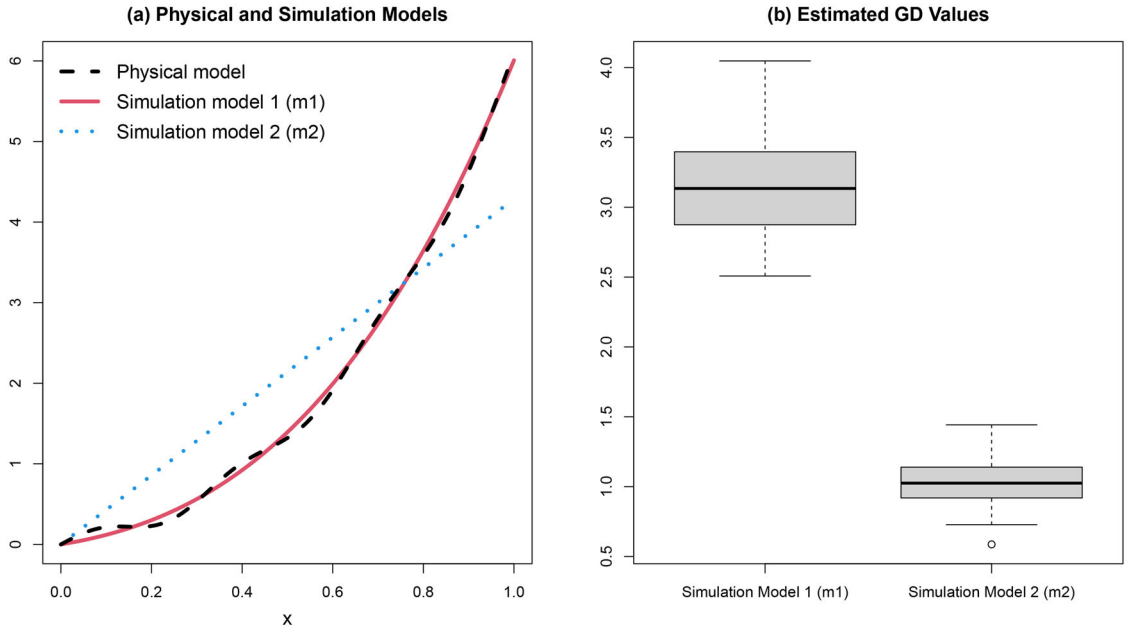

Figure 2. (a) The physical model and two simulators. (b) The estimates of the generalized degrees of freedom for the two simulators.

Table 3. The range and description of the input variables in the T-cell adhesion frequency assay experiments.

Type of variables		Physical experiments	Simulators		Description (*s* represents second)	Range
			CC	RP		
Control variables	x_{wt}	✓	✓	✓	Waiting time in between contacts (s)	[1, 6]
	x_{ct}	✓	✓	✓	Cell–cell contact time (s)	[0.25, 5]
Calibration parameters	x_{Kc}		✓		Kinetic proofreading rate for activation of cluster (1/s)	[0.1, 100]
	x_{Kf}		✓	✓	On-rate enhancement of inactive TCRs ($\mu m^2/s$)	$[10^{-8}, 10^{-10}]$
	x_{Kr}		✓	✓	Off-rate enhancement of inactive TCRs (1/s)	[0.1, 10]
	$x_{r,p}$		✓	✓	Off-rate enhancement of activated TCRs (1/s)	[0.01, 100]

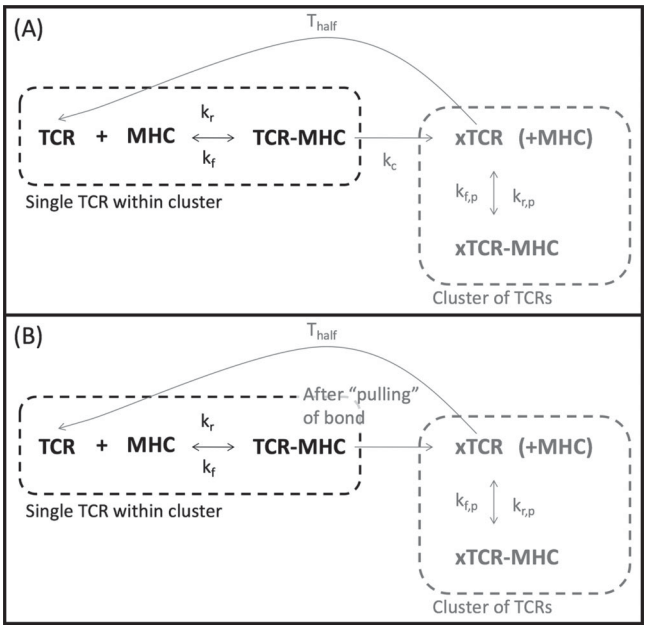

Figure 3. Two simulators capturing two biological mechanisms.

or involve force due to the pulling of the TCR-pMHC bond (Rittase 2018). Biologists are interested in understanding which mechanism is behind the recognition process, but it cannot be directly detected by physical experiments. Therefore, the goal of this study is to identify the optimal mechanism based on the observed experimental data from the laboratory.

Two control variables, contact time x_{ct} and waiting time x_{wt}, are involved both in the lab experiments and in the simulators. Denote $\mathbf{x} = (x_{wt}, x_{ct})$. Four calibration parameters, denoted by $x_{Kf}, x_{Kr}, x_{Kr,p}$, and x_{Kc}, are involved in the CC mechanism, while only the first three of them are involved in the RP mechanism. The descriptions for the variables are given in Table 3, and further details can be found in Sung et al. (2020a). The two mechanisms are simulated by the Gillespie (1976) algorithm, which is a stochastic simulation algorithm. The experimental outputs are binary, indicating a TCR-pHMC binding or not. A 60-run OA-based Latin hypercube design (Tang 1993) is implemented for the two simulators, and each design consists of 10 replicates to capture the cell–cell variability. Therefore, the sample size of the computer experiment is 600 for each mechanism. For the physical experiments, the sample size is $n = 272$ and the settings of x_{ct} and x_{wt} are randomly chosen from the sample space $[0.25, 5] \times [1, 6]$.

Table 4. The leave-one-out cross-validation errors and the estimated degrees of freedom for the two simulators.

Simulator	LOOCV	Generalized degrees of freedom
CC mechanism	0.102	5.358
RP mechanism	0.146	4.950

Given the binary binding outcomes $y(\mathbf{x})$ observed in the laboratory, the true process is defined as the binding probability, $\xi(\mathbf{x}) = P(y(\mathbf{x}) = 1)$, and estimated by a kernel logistic regression

$$\text{logit}\{\hat{\xi}(\mathbf{x})\} = \hat{\alpha}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \Psi(\mathbf{x}_i, \mathbf{x}), \tag{14}$$

where $\text{logit}\{\cdot\}$ is the logistic link function, $\{\hat{\alpha}_i\}_{i=0}^n$ are the estimated coefficients, and $\Psi(\mathbf{x}', \mathbf{x})$ is the Matérn kernel with roughness parameter $\nu_0 = 2.5$. Define $p(\mathbf{x}; \boldsymbol{\theta}) = P(y^s(\mathbf{x}; \boldsymbol{\theta}) = 1)$, where $y^s(\mathbf{x}; \boldsymbol{\theta})$ is the simulated binary outcomes, and its emulators are constructed by the generalized Gaussian process models (Sung et al. 2020b)

$$f_k(\mathbf{x}; \boldsymbol{\theta}_k, \boldsymbol{\beta}_k) = \text{logit}\{p(\mathbf{x}; \boldsymbol{\theta}_k)\} \sim GP(\mu_k, \sigma_k^2 \Phi((\mathbf{x}', \boldsymbol{\theta}_k'), (\mathbf{x}, \boldsymbol{\theta}_k))), \tag{15}$$

where $\Phi((\mathbf{x}', \boldsymbol{\theta}_k'), (\mathbf{x}, \boldsymbol{\theta}_k))$ is the Matérn kernel with roughness parameter $\nu = 1.5$, $\boldsymbol{\beta}_k = (\mu_k, \sigma_k^2, \rho_k)$ includes the mean, variance, and range parameters for CC simulator ($k = 1$) or RP simulator ($k = 2$). $\boldsymbol{\beta}_k$ is estimated by empirical maximum likelihood method as in Section 4. The calibration parameters are estimated by minimizing the L_2 discrepancy proposed by Sung et al. (2020a).

The leave-one-out cross-validation errors for the two simulators are summarized in Table 4 along with the estimated generalized degrees of freedom. The optimal simulator is the CC mechanism because its LOOCV is smaller, while its sensitivity is slightly higher than that for the RP mechanism. From a biological perspective, the selection of the CC mechanism indicates that the molecules have conformational changes due to the TCR-pMHC binding. Analyzing the CC mechanism using all the data, we have $\hat{\boldsymbol{\beta}}_1 = (\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\rho}_1) = (-0.322, 1.732, 3.089)$, and $\hat{\boldsymbol{\theta}} = (1.560, 8.563 \times 10^{-7}, 1.425, 1.589)$. By plugging in the estimated calibration parameters, the simulated binding probability according to the CC mechanism (red dashed lines) in

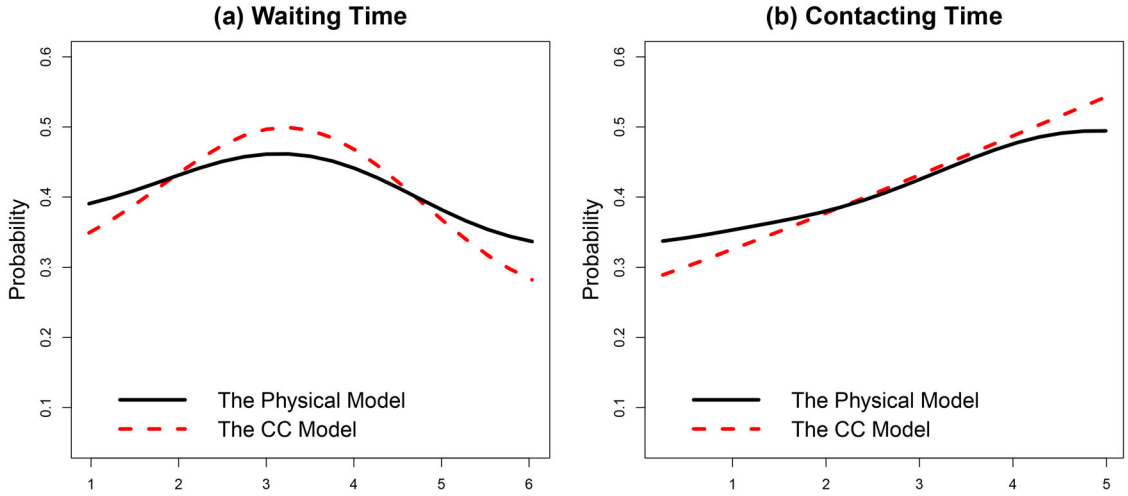

Figure 4. The fitted adhesion models from the physical experiment and from the computer experiment of the CC model for two control variables: waiting time and contacting time.

Figure 4 as a function of the two control variables, waiting time and contact time. It appears that the selected optimal simulator, CC mechanism, can reasonably capture the trend observed in the lab experiments.

6. Summary and Concluding Remarks

In many applications, identifying an optimal simulator for the observed physical experiments can provide scientific insights that are not available from lab experiments. There is, however, no systematic statistical method to tackle this problem. We propose a new criterion based on the idea of leave-one-out cross-validation. Theoretical properties of the selection method based on the criterion and the estimated optimal simulator are discussed. It is also shown that asymptotically the proposed approach includes the L_2 calibration method as a special case. Simulation studies are conducted to demonstrate the performance of the proposed method. By applying the proposed method, the selected optimal T-cell signaling simulator suggests that the true binding mechanism is through conformational changes in molecules, which may shed new light on the antigen recognition mechanism in human immune system.

As pointed out by one of the reviewers, an important and interesting research is to understand the convergence properties of the estimated optimal emulator f_T, such as the convergence rate to the optimal emulator f_0. A promising direction is to extend recent results in Wang, Tuo, and Wu (2020), which is for deterministic functions, to stochastic functions. This work will be considered in a future research.

Supplementary Materials

The online supplemental material contains more technical details of this paper, including the assumptions used in Theorems 3.3 and 3.4, and the detailed proofs of the lemma, proposition, and theorems.

Acknowledgment

The authors acknowledge the support from the National Science Foundation (DMS 1660504, 1914632, and 1660477).

Funding

Wu's work is supported by NSF grants DMS 1660504 and 1914632, and Hung's work is supported by NSF grants DMS 1660477.

References

Bastos, L. S., and O'Hagan, A. (2009), "Diagnostics for Gaussian Process Emulators," *Technometrics*, 51, 425–438. [1264]

Brauer, F., and Castillo-Chavez, C. (2012), *Mathematical Models in Population Biology and Epidemiology*, New York: Springer. [1264]

Burnham, K. P., and Anderson, D. R. (2011), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer. [1264]

Efron, B. (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470. [1265,1266]

Fang, K.-T., Li, R., and Sudjianto, A. (2006), *Design and Modeling for Computer Experiments*, Boca Raton, FL: Chapman & Hall/CRC. [1264,1265]

Gillespie, D. T. (1976), "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *Journal of Computational Physics*, 22, 403–434. [1269]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [1265]

Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., and Rutter, E. (2013), "Prediction and Computer Model Calibration Using Outputs From Multifidelity Simulators," *Technometrics*, 55, 501–512. [1267]

Joseph, V. R., Gul, E., and Ba, S. (2015), "Maximum Projection Designs for Computer Experiments," *Biometrika*, 102, 371–380. [1267]

Kennedy, M. C., and O'Hagan, A. (2000), "Predicting the Output From a Complex Computer Code When Fast Approximations Are Available," *Biometrika*, 87, 1–13. [1264]

Lakshmi, C., and Iyer, S. A. (2013), "Application of Queueing Theory in Health Care: A Literature Review," *Operations Research for Health Care*, 2, 25–39. [1264]

Overstall, A. M., and Woods, D. C. (2016), "Multivariate Emulation of Computer Simulators: Model Selection and Diagnostics With Application to a Humanitarian Relief Model," *Journal of the Royal Statistical Society*, Series C, 65, 483–505. [1264]

Richardson, L. F. (2007), *Weather Prediction by Numerical Process*, Cambridge: Cambridge University Press. [1264]

Rittase, W. R. (2018), "Combined Experimental and Modeling Studies Reveal New Mechanisms in T cell Antigen Recognition," PhD thesis, Georgia Institute of Technology. [1268,1269]

Santner, T. J., Williams, B. J., Notz, W., and Williams, B. J. (2018), *The Design and Analysis of Computer Experiments* (2nd ed.), New York: Springer. [1264,1265,1267]

Shawe-Taylor, J., and Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge: Cambridge University Press. [1265]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [1265]

Sung, C.-L., Hung, Y., Rittase, W., Zhu, C., and Wu, C. F. J. (2020a), "Calibration for Computer Experiments With Binary Responses and Application to Cell Adhesion Study," *Journal of the American Statistical Association*, 115, 1664–1674. [1265,1266,1269]

——— (2020b), "A Generalized Gaussian Process Model for Computer Experiments With Binary Time Series," *Journal of the American Statistical Association*, 115, 945–956. [1269]

Tang, B. (1993), "Orthogonal Array-Based Latin Hypercubes," *Journal of the American Statistical Association*, 88, 1392–1397. [1269]

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, New York: Springer. [1265]

Tuo, R., and Wu, C. F. J. (2015), "Efficient Calibration for Imperfect Computer Models," *Annals of Statistics*, 43, 2331–2352. [1265,1266,1267]

Tuo, R., Wu, C. F. J., and Yu, D. (2014), "Surrogate Modeling of Computer Experiments With Different Mesh Densities," *Technometrics*, 56, 372–380. [1264]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM. [1265]

Wang, W., Tuo, R., and Wu, C. F. J. (2020), "On Prediction Properties of Kriging: Uniform Error Bounds and Robustness," *Journal of the American Statistical Association*, 115, 920–930. [1270]

Wong, R., Storlie, C., and Lee, T. (2017), "A Frequentist Approach to Computer Model Calibration," *Journal of the Royal Statistical Society*, Series B, 79, 635–648. [1265]

Wood, S. N., Pya, N., and Säfken, B. (2016), "Smoothing Parameter and Model Selection for General Smooth Models," *Journal of the American Statistical Association*, 111, 1548–1563. [1264]

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131. [1266]