Calibrations and validations of biological models with an application on the renal fibrosis

Georgios Karagiannis
Department of Mathematical Sciences, University of Durham, UK
georgios.karagiannis@durham.ac.uk

Wenrui Hao
Department of Mathematics, Penn State University, College Station, PS, USA
wxh64@psu.edu

Guang Lin
Department of Mathematics, School of Mechanical Engineering,
and Department of Statistics (by Courtesy), Purdue University, West Lafayette, IN, USA
lin491@purdue.edu

January 13, 2021

Abstract

We calibrate a mathematical model of renal tubulointerstitial fibrosis by Hao et al. [5], which is used to explore potential drugs for Lupus Nephritis, against a real data-set of 84 patients. For this purpose, we present a general calibration procedure which can be used for the calibration analysis of other biological systems as well. Central to the procedure is the idea of designing a Bayesian Gaussian process (GP) emulator that can be used as a surrogate of the fibrosis mathematical model which is computationally expensive to run massively at every input value. The procedure relies on detecting influential model parameters by a GP based sensitivity analysis, and calibrating them by specifying a maximum likelihood criterion, tailored to the application, which is optimized via Bayesian global optimization.

Keywords: Renal tubulointerstitial fibrosis, calibration, surrogate model, Gaussian process emulation

1 Introduction

An autoimmune disease, such as Lupus Nephritis, starts with inflammation in the kidney, which may lead to renal tubulointerstitial fibrosis. The renal fibrosis is characterized by abnormal excessive deposition of fibrous proteins, such as collagen, and is often progressive leading to kidney dysfunction or failure. A mathematical model of renal tubulointerstitial fibrosis was recently developed by Hao et al.[5] and was used to explore

potential drugs. This mathematical model is based on the network of cells and cytokines of renal fibrosis progression which is represented by a system of partial differential equations (PDEs) in a section of the kidney tissue. More recently Hao et al. [4] extended this model to IPF and used it to explore the efficacy of drugs. These computational models enable predictions of fibrosis progression and can provide a systematic way to quantify the precision medicine for fibrosis patients. However, developing these models depends on the knowing biology behind the fibrosis progressions but lacks of unknown qualities. Then further model refinement needs a calibration technique for the model validation and parameter analysis that aligns model and fibrosis progression.

In this paper, we present a computational method for calibrating the renal tubulointerstitial fibrosis mathematical model in Hao et al. [5] against real data. This method is based on Bayesian global optimization, global sensitivity analysis, and the Gaussian process emulation. It is a general approach and can be easily extended to other biological models for a deeper consideration of calibration to demonstrate the data-driven modeling purpose. Here, as fibrosis model is expensive to run massively, the Gaussian process (GP) emulator [17] is utilized as a computationally cheap probabilistic surrogate model aiming at representing computationally expensive functions and quantifying uncertainties due to approximations. Initially, the influential model parameters are identified via a surrogate model based global sensitivity analysis [11] as a part of a screening procedure to reduce the computational requirements during the calibration procedure. Optimal values for the model parameters are discovered by minimizing a maximum likelihood criterion, measuring the proximity of the real data and the model output, via a Bayesian global optimization. The criterion is specified properly in order to take into account differences between the three patient groups (low, intermediate, and high fibrosis).

The layout of the paper is as follows: we describe the mathematical model simulating Fibrosis in Section 2, we present the statistical methods that we use to analyze the application in Section 3, we perform the analysis of the fibrosis mathematical model and discuss the results in Section 4, we conclude in Section 5.

2 Mathematical model of the renal fibrosis

The model developed in Hao et al [5] is used to describe the casual network of the renal fibrosis which starts the injured glomerular compartment that communicates with the tubulointerstitial compartment. By responding to glomerular immune complexes, monocytes enter glomeruli from the circulation differentiate into tissue macrophages and then escape from damaged glomeruli into the tubulointerstitium. Similarly, soluble pro-inflammatory factors could “leak" out of damaged glomeruli and activate tubular epithelial cells (TECs). Activated TECs secrete some pro-inflammatory mediators including chemoattractants like monocyte chemotactic protein-1 (MCP-1), that recruit additional circulating monocytes to the tubulointerstitial space. Infiltrating monocytes become activated macrophages, with activated TECs, facilitate the progression from inflammation to interstitial fibrosis by secreting several cytokines and growth factors such as platelet-derived growth factor (PDGF), transforming growth factor-beta (TGF-β), matrix metalloproteinase (MMP) and tissue inhibitor of metalloproteinase (TIMP) (See more details in Figure 1). All these cytokines are involved in the regulation of tissue fibrosis: TGF-β, along with TEC-derived basic fibroblast growth factor (bFGF) increases the proliferation of interstitial fibroblasts; PDGF and TGF-β transform fibroblasts to myofibroblasts, which produce extracellular matrix (ECM). The imbalance between MMP and its inhibitor

TIMP facilitates the accumulation of ECM and formation of interstitial fibrosis.

In order to model different cells dispersion and cytokines' diffusion (due to the renal fibrosis progression) shown in Fig. 1, a system of convection-diffusion equations has been used in Hao et al. [5]. The equation for each species C_i $(1 \leqslant i \leqslant N)$ has the following form [5]:

$$\frac{\partial C_i}{\partial t} - D_{C_i} \Delta C_i = F_{C_i}(\mathbf{C}) \text{ in } \Omega \tag{1}$$

where D_{C_i} is the diffusion/dispersion coefficient and $F_{C_i}(\mathbf{C})$ is a function which models all the species' interactions in the molecular networks shown in Figure 1. We will not write $F_{C_i}(\mathbf{C})$ explicitly here for each C_i (see Hao et al. [4, 5] for more details). The system of diffusion equations, Eq. (1), is solved by the central difference scheme of finite difference method in discretizing the space coupled with the forward Euler method in discretizing the time.

This model has been compared with the patient data qualitative [5]. However it did not calibrate the model and perform the parameter sensitivity analysis which is very important to illustrate the usefulness of the model in clinical practice. In the present paper, we will further calibrate this model with respect to the available patient data.

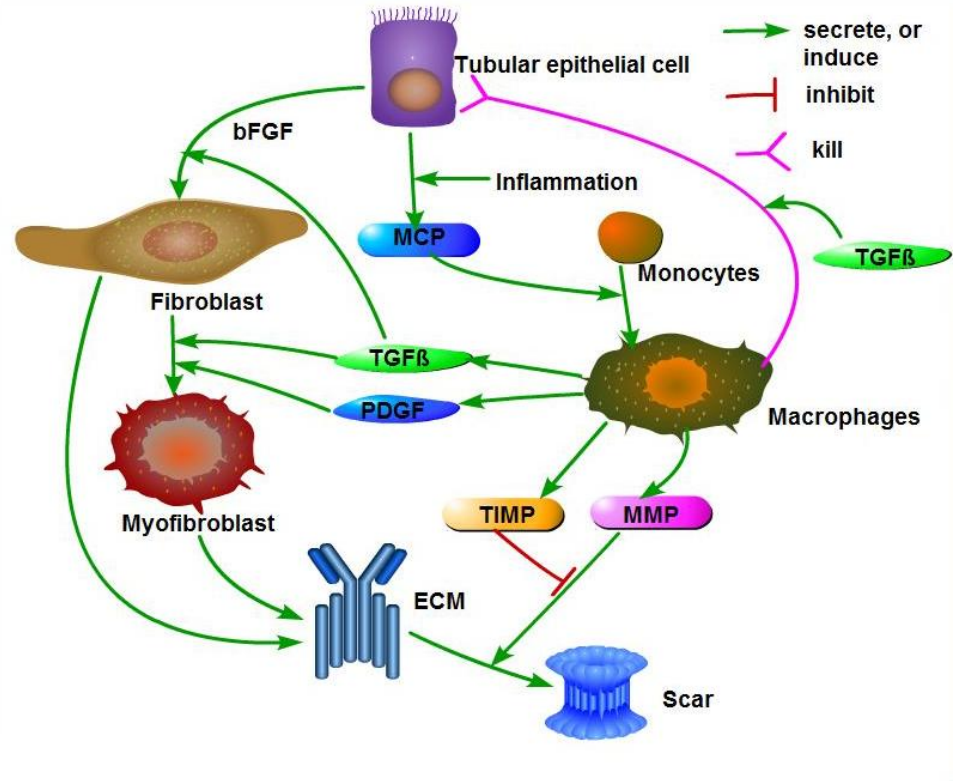


Figure 1: Schematic networks of the renal fibrosis.

3 Model calibration statistical methods

We present a two stage procedure for the calibration of the fibrosis mathematical model (1) which involves sensitivity analysis and Bayesian global optimization.

3.1 Gaussian process emulators of unknown functions

Emulators of computationally expensive or unknown functions are cheap probabilistic surrogate models [17, 18, 19] aiming at quantifying uncertainties due to approximations. Assume there is interest in representing a function whose analytic expression is not directly available; e.g., the output of a simulator. Let the unknown function be $f(\cdot)$, that inquires a set of inputs $x := (x_1, ..., x_d) \in \mathcal{X} \subset \mathbb{R}^d$ and produces an output $y = f(x) \in \mathbb{R}$.

Central to the Gaussian process (GP) emulation [17, 23] is the idea of modeling $f(\cdot)$ as a Gaussian process which can be fully specified by its mean function and correlation function. Here, we specify

$$f(\cdot)|\beta, \psi \sim \mathrm{GP}(\mu_0(\cdot|\beta), \sigma^2 r_0(\cdot, \cdot|\psi)),$$

where β and ψ are unknown hyper-parameters of the mean $\mu_0(\cdot|\beta)$ and correlation $r_0(\cdot, \cdot|\psi)$ functions, and σ^2 is an unknown scale parameter controlling the variance at each input that need to be specified accordingly. Although $f(\cdot)$ is unknown, it is often possible to exist a priori knowledge regarding some properties of the $f(\cdot)$, e.g. smoothness; such information can be incorporated through the specification of $\mu_0(\cdot|\beta)$ and $r_0(\cdot, \cdot|\psi)$. The mean function $\mu_0(\cdot|\beta)$ traditionally is modeled as

$$\mu_0(x|\beta) = h^\top(x)\beta,$$

where $h(\cdot) := (h_1(\cdot), ..., h_q(\cdot))$ are known basis functions, e.g. polynomial bases [26], and $\beta \in \mathbb{R}^q$ are unknown coefficients, because it can satisfactory represent long scale dependences. The covariance function $r_0(\cdot, \cdot|\psi)$ must be specified such that every covariance matrix with elements $\{r_0(x_i, x_j|\psi)\}$ is non-negative defined. In the literature, there are several covariance function families with different properties [18]. A commonly used class of covariance functions is the Matern [23] because of its flexibility. Specifically, a correlation function based on the ν-Matern kernel is such $r_0(x, x'|\psi) = k_{\mathrm{Matern},\nu}(\sum_{j=1}^d (x_j - x'_j)^2 \psi_j^{-1})$, where

$$k_{\mathrm{Matern},\nu}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}d)^\nu K_\nu(\sqrt{2\nu}d),$$

$\{\psi_j^{-1}\}$ are positive length-scale parameters describing how smooth $f(\cdot)$ is with respect to the j-th input, $K_\nu(\cdot)$ is a modified Bessel function [1], and $\nu > 0$ is a smoothness parameter that describes the differentiability of the function. The use of a 'nugget' term $\kappa > 0$ in the covariance function, as $r_0(x, x'|\psi) = k_{\mathrm{Matern},\nu}(\sum_{j=1}^d (x_j - x'_j)^2/\psi_j) + \kappa$ can lead to a better statistical properties and computational stability [3]; here we consider κ as a fixed value specified by the researcher.

We assume that there is available a (training) data-set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ of size n that contains realizations $y_i := f(x_i)$ at design points x_i for $i = 1, ..., n$. The joint distribution of the outputs y (forming the likelihood function) is a multivariate Normal distribution

$$y|\beta, \psi, \sigma^2 \sim \mathrm{N}_n(H\beta, \sigma^2 R(\psi)),$$

where σ^2 is an unknown scalar parameter, R is the correlation matrix such that $[R(\psi)]_{i,j} = r_0(x_i, x_j|\psi)$, and H is a design matrix such that $H = (h(x_1), ..., h(x_n))^\top$. Working in the Bayesian framework, we assign weak prior $\pi(\beta, \sigma^2) \propto 1/\sigma^2$ on the unknown parameters (β, σ^2), and a user specified proper prior $\pi(\psi)$ on ψ.

The predictive distribution of $f(\cdot)$ at new input points x' and x'' given the observed y results by integrating out (β, σ^2) from the joint distribution $[f(x''), f(x'), y, \beta, \sigma^2, \psi]$ and conditioning on y, ψ, as a Student-t process [20, 18, 24]:

$$f(\cdot)|y, \psi \sim \mathrm{StP}(\mu_n(\cdot|\psi), v_n(\psi)C_n(\cdot, \cdot|\psi), n - q), \tag{2}$$

with degrees of freedom $n - q$, mean and shape function

$$\begin{aligned}
\mu_n(x|\psi) =& h(x)^\top \hat{\beta}_n(\psi) + r(x|\psi)^\top R^{-1}(\psi)(y - H\hat{\beta}_n(\psi)) && (3)\\
C_n(x, x'|\psi) =& [r_0(x, x'|\psi) - r(x|\psi)^\top R^{-1}(\psi) r(x'|\psi) \\
&+ (h(x) - r(x|\psi)^\top R^{-1}(\psi)H)(H^\top R^{-1}(\psi)H)^{-1}](h(x') - r(x'|\psi)^\top R^{-1}(\psi)H)^\top, && (4)\\
v_n(\psi) =& \frac{y^\top (R^{-1}(\psi) - R^{-1}(\psi)H(HR^{-1}(\psi)H^\top)^{-1}H^\top R^{-1}(\psi))y}{n - q} && (5)
\end{aligned}$$

where $r(\cdot|\psi) := (r_0(\cdot, x_1|\psi), ..., r_0(\cdot, x_n|\psi))$, and $\hat{\beta}_n(\psi) = (HR^{-1}(\psi)H^\top)^{-1}H^\top R^{-1}(\psi)y$. The predictive distribution (2) can be used as an emulator for $f(\cdot)$, while the posterior mean $\mu_n(x|\psi)$ can be used as a surrogate model for $f(\cdot)$, provided we know the unknown ψ.

To learn the unknown ψ, we consider the Maximum A posteriori (MAP) estimation which requires one to maximize the log marginal posterior distribution density of ψ

$$\pi(\psi|y) \propto \mathrm{N}_n(y|H\hat{\beta}_n(\psi), \sigma^2 R(\psi))\pi(\psi),$$

where $\mathrm{N}_d(y|b, B)$ denotes the density of the d-dimensional Normal distribution with mean b and covariance B. Namely, we seek the MAP estimates of ψ as

$$\hat{\psi} = \arg\min_{\psi} \{\log(\pi(\psi)) - \frac{1}{2}\log\det(R(\psi)) - \frac{1}{2}\log\det(H^\top R^{-1}(\psi)H) - (v_n(\psi))^{\frac{n-q}{2}}\}, \tag{6}$$

This is essentially an Empirical Bayes approach which can be seen as a fast approximation of the fully Bayesian method. To solve (6), one can implement any optimization algorithm, here we use the Parallel and Interacting Stochastic Approximation Annealing (PISAA) algorithm [9].

3.2 Surrogate based global sensitivity analysis

Sensitivity analysis (SA) quantifies how uncertainties in the output of the simulator can be apportioned to uncertainties in the inputs throughout the input space. It can be used to identify non-influential parameters as a part of a screening procedure in order to reduce model complexity. This can mitigate the computational cost in a subsequent model calibration due to high dimensionality of the input space.

According to Hoeffding-Sobol expansion [6], any function $f(\cdot)$ can be decomposed into summands of

increasing dimensionality as

$$f(x) = f_o + \sum_{i=1}^{d} f_i(x_i) + \sum_{1 \leqslant i < j \leqslant d} f_{i,j}(x_i, x_j) + ... = \sum_{\boldsymbol{i} \in \mathcal{I}} f_{\boldsymbol{i}}(x_{\boldsymbol{i}}),$$

under certain orthogonality conditions [21], where $\mathcal{I}$ is a collection of all possible index subsets. This allows the total variance $D = \mathrm{Var}(f(x))$ of $f(x)$ to be factorized as

$$D = D_o + \sum_{i=1}^{d} D_i + \sum_{1 \leqslant i < j \leqslant d} D_{i,j} + ... = \sum_{\boldsymbol{i} \in \mathcal{I}} D_{\boldsymbol{i}},$$

where $D_{\boldsymbol{i}} = \mathrm{Var}(\mathrm{E}(f(x)|x_{\boldsymbol{i}})) - \sum_{\boldsymbol{j} \subset \boldsymbol{i}} \mathrm{Var}(\mathrm{E}(f(x)|x_{\boldsymbol{j}}))$ are partial variances [22]. The Sobol sensitivity index for $\boldsymbol{i}$ is defined as

$$S_{\boldsymbol{i}} = \frac{D_{\boldsymbol{i}}}{D}. \tag{7}$$

For instance, $S_i = D_i/D$ is the first-order sensitivity index for input parameter x_i and measures its main effect on the output, while $S_{i,j} = D_{i,j}/D$ is the second-order sensitivity index that measures the part of the variance explained by x_i and x_j jointly but not individually. Often it is computationally expensive to evaluate all possible $\{S_{\boldsymbol{i}}; \forall \boldsymbol{i} \in \mathcal{I}\}$, and hence the main interest focuses on the total sensitivity indices $S_{T_i} = S_i + \sum_{\{\boldsymbol{i}: i \in \boldsymbol{i}\}} S_{\boldsymbol{i}}$ which measures the total effect of i-th input parameter on the output. Small values of $S_{T_i} \approx 0$ imply that the i-th input parameter is non-influential and hence can be fixed in the subsequent model calibration.

Evaluation of $S_{\boldsymbol{i}}$ in (7) cannot be performed analytically due to the intractable integrals. To address this issue, approximation modified versions of (7) such as the Monte Carlo based methods [2, 7, 14], meta-model based methods [27, 16, 11], or screening based methods [15] have been developed. Here, we focus on the GP based Sobol sensitivity analysis [11] according to which the function $f(\cdot)$ is substituted by the surrogate model (2) and then Sobol indices are estimated via Monte-Carlo integration. The estimator of (7) is

$$\tilde{S}_{\boldsymbol{i}}^{(n,m)} = \frac{\tilde{D}_{\boldsymbol{i}}}{\tilde{D}} = \frac{\frac{1}{m}\sum_{t=1}^{m} \tilde{f}(x_i)\tilde{f}(x_{-\boldsymbol{i}}) - (\frac{1}{m}\sum_{t=1}^{m} \tilde{f}(x_i))(\frac{1}{m}\sum_{t=1}^{m} \tilde{f}(x_{-\boldsymbol{i}}))}{\frac{1}{m}\sum_{t=1}^{m} \tilde{f}(x_i)^2 - (\frac{1}{m}\sum_{t=1}^{m} \tilde{f}(x_i))^2} \tag{8}$$

where $\tilde{f}(\cdot) = \mu_n(x|\hat{\psi})$ denotes the surrogate model (3), n is the data-set size required to train the surrogate model, and m is the sample size for the Monte Carlo estimates in (8). The evaluation of the distribution of $\tilde{S}_{\boldsymbol{i}}^{(n,m)}$ (as well as the estimation of its statistics mean, and standard error) is performed numerically via Bootstrap sub-sampling algorithm [11] which takes into account both uncertainty of the GP emulator meta-model and Monte Carlo integration error.

3.3 Bayesian global optimization

We consider the problem of optimizing an objective function $f(\cdot)$ of p input parameters $x := (x_1, ..., x_p)$. Namely we need to discover input value (location) x^* and associated $f(x^*)$ such as

$$x^* = \arg \min_{x \in \mathcal{X} \subset \mathbb{R}^p} f(x).$$

BGO aims at optimizing objective functions which are extremely computationally expensive to be directly evaluated excessively many times and whose analytic expression and derivatives may not be available. Central to BGO is the idea of building a probabilistic surrogate model for the objective function and using it to define a utility function called 'information acquisition function' (IAF) whose role is to guide the search for the optimum value.

The IAF is defined such that high acquisition corresponds to potentially high values of the objective function either because the prediction is high or the uncertainty is large. The expected improvement (EI) [8] is perhaps the most commonly used IAF because it accounts for both exploration (sampling from areas of high uncertainty) and exploitation (sampling from areas likely around the global minimum) in an efficient and tractable manner. The improvement is defined as $\mathrm{I}(x) = \max(0, f(x^*) - f(x))$, where x^* is the currently best location discovered. Then the expected improvement $\mathrm{EI}(x)$ with respect to the process (2) is

$$\mathrm{EI}(x|\hat{\psi}) = \begin{cases} \sigma_n^2(x|\hat{\psi}) z_n(x|\hat{\psi}) \Phi_{\mathrm{St}}(z_n(x|\hat{\psi})) + \frac{n-q}{n-q-1}\left(1 + \frac{z_n(x|\hat{\psi})^2}{n-q}\right) \sigma_n(x|\hat{\psi}) \phi_{\mathrm{St}}(z_n(x|\hat{\psi})), & \text{if } \sigma_n(x|\hat{\psi}) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $z(x|\hat{\psi}) = \frac{f(x^*) - \mu_n(x|\hat{\psi})}{\sigma_n(x|\hat{\psi})}$, $\hat{\psi}$ is the MAP in (6), while $\mu_n(x|\hat{\psi})$ and $\sigma_n^2(x|\hat{\psi}) = v_n(\hat{\psi}) C_n(x, x|\hat{\psi})$ denote the predictive mean (3) and variance (4, 5) of the GP emulator. Here, $\Phi_{\mathrm{St}}(\cdot)$ and $\phi_{\mathrm{St}}(\cdot)$ denote the standard Student-t cumulative and probability density function. We remark that the EI in (9) is the expectation of the improvement $I(x)$ with respect to a Student-t process (2) and hence it is different that the standard BGO where the expectation is with respect to a Gaussian process. The present treatment allows the variance component $\sigma_n^2(x|\hat{\psi})$ in (9) to depend not only on the locations x but also on $f(x)$; this can improve the exploration ability of BGO to detect more useful input locations.

The value of $\mathrm{EI}(x|\hat{\psi})$ increases at a given x when either $\mu_n(x|\hat{\psi})$ is smaller than the currently best value $f(x^*)$ discovered or there is a large amount of uncertainty in the predictor $\mu_n(x|\hat{\psi})$ when $\sigma_n(x|\hat{\psi})$ is large. It is worth mentioning that although evaluating and optimizing $f(\cdot)$ directly can be computationally expensive, computing and optimizing EI (9) is cheap. EI can be optimized with an existing optimization algorithm, here we use the PISAA [9].

The BGO procedure initiates by collecting an initial data-set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ by directly evaluating the objective function at n input location values selected via a space filling method such as Latin hyper-cube sampling (LHS) [12]. It proceeds by performing a recursion: evaluate the GP emulator (2) and the IE function (9) given the current data-set; discover input values with the largest acquisition information (IE value); and append the objective function values evaluated at these input values to the the data-set D. The recursion iterates until a termination criterion is met. A termination criterion can be used based on the maximum number of iterations performed, or the convergence rate of the objective function such as $|f^* - f'| < \epsilon_{\mathrm{tol.}}|f^*|$ where f' is the new best value and $\epsilon_{\mathrm{tol.}}$is a tolerance. The procedure is presented in Figure 2.

Initialize

Generate an initial training data-set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, by selecting n input points $\{x_i\}$ via a LHS and computing directly the output of the objective function $\{y_i = f(x_i)\}$.

Iterate

1. Given current data-set D, train the GP emulator of $f(x)$, find $\hat{\psi}$ and calculate the expected improvement $\mathrm{EI}(x|\hat{\psi})$.
2. Find the input values x' that maximize the expected improvement by solving $x' = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x|\hat{\psi})$.
3. Augment the data-set as $D \leftarrow D \cup \{(x', y')\}$ by computing directly $y' = f(x')$ at the input values x'.
4. Stop if the termination criterion $|f - f'| < \epsilon_{\mathrm{tol.}}|f|$ is satisfied, where f is the current best value.

Figure 2: Work-flow of the Bayesian global optimization algorithm

4 Analysis

We analyze the renal tubulointerstitial fibrosis mathematical model (simulator) in (1). The model aims at predicting the uMCP1 ($ng/mgcr$) at three levels of fibrosis severity (low, intermediate, and high) given a set of input (model) parameters λ_{ET}, λ_{TM}, λ_{GM}, λ_{QM}, λ_{QrM}, λ_{PM}, $\lambda_{\rho\mathrm{f}}$, $\lambda_{\rho\mathrm{m}}$, $\lambda_{\rho\mathrm{T}}$, λ_{fE}, λ_{mfT}, and λ_{mfG}. The default values and the ranges of the input parameters are presented in Table 1.

Parameter	Default	Range	Parameter	Default	Range
λ_{ET}	2	$[0.2, 3.8]$	$\lambda_{\rho\mathrm{f}}$	$3e-3$	$[0.0003, 0.0057]$
λ_{TM}	$3e-2$	$[0.003, 0.057]$	$\lambda_{\rho\mathrm{m}}$	$6e-3$	$[0.0006, 0.0114]$
λ_{GM}	$2.4e-5$	$[2.4e-6, 4.56e-05]$	$\lambda_{\rho\mathrm{T}}$	1	$[0.1, 1.9]$
λ_{QM}	$3e-4$	$[3e-5, 0.00057]$	λ_{fE}	$1.05e-2$	$[1.05e-3, 0.01995]$
λ_{QrM}	$6e-5$	$[6e-6, 0.000114]$	λ_{mfT}	$1.2e-1$	$[0.012, 0.228]$
λ_{PM}	$3e-3$	$[3e-4, 0.0057]$	λ_{mfG}	$1.2e-1$	$[0.012, 0.228]$

Table 1: Default values, and ranges of the model parameters

We are interested in identifying which of these model parameters significantly influence the output of the simulator –and hence its behavior– first, and then calibrating these parameters against a real data-set. For the former, we will perform sensitivity analysis, while for the later we will perform Bayesian global optimization to minimize a maximum likelihood criterion suitable for the application.

Sensitivity analysis The mathematical model for fibrosis has 12 input parameters defined in ranges presented in Table 1. We wish to determine non-influential model parameters which can be fixed in the subsequent calibration procedure. We perform GP based global sensitivity analysis described in Section 3.2, which estimates the Sobol indices by substituting the output function of the simulator via surrogate model and using Monte-Carlo integration. This approach is suitable for our application because the simulator is very expensive to run directly and because it has a large number of input parameters.

We are mainly interested in the total sensitivity Sobol indices. To build the GP emulator, we considered the prior mean function to be constant and the covariance function to be a tensored 5/2-Matern kernel. The emulator was trained via MAP, against a training data-set of size 1000 where the input points were selected via LHS. The Monte Carlo estimator of Sobol indices was defined on a sample with size 5000.

The estimates of the total sensitivity Sobol indices are presented in Figure 3, for each fibrosis sensitivity level. At low and intermediate fibrosis severity levels, we observe that the input parameter λ_{PM} significantly explains most of the variability of the simulator output (with Sobol indexes equal to 99% and 98% respectively). At the high fibrosis severity level, we observe that the input parameters λ_{PM} and λ_{ET} can explain most of the variability of the simulator output (with Sobol indexes equal to 85% and 12% respectively). The rest input parameters are non-influential as the confidence intervals of their corresponding Sobol indexes include 0% values. Therefore, in the subsequent calibration of the simulator under consideration, interest lies in calibrating the influential model parameters λ_{ET} and λ_{PM} while keeping the rest fixed and equal to the default parameters in Table 1.

Calibration analysis We wish to calibrate the model parameters of the fibrosis mathematical model (1), given that the rest input parameters are fixed and equal to the default values in Table 1. In order to calibrate the model, we use the patient data provided in Hao et al. [5]. Urine was collected at the time of diagnostic kidney biopsy of Systemic lupus erythematosus (SLE) patients who developed clinical signs of kidney involvement. Urine MCP-1 was measured by specific ELISA; Urine TGF-β was also measured by specific ELISA obtained from R & D systems. The simulator is calibrated against observed data (uMCP1 ($ng/mgcr$) measurements) collected from 84 individuals. The individuals where categorized according to their fibrosis severity in 3 groups: 'low', 'intermediate', and 'high' severity level.

To calibrate the simulator with respect to model parameters λ_{ET} and λ_{PM}, we seek optimal values λ^*_{ET}, λ^*_{PM} for λ_{ET} and λ_{PM}, such that the fibrosis model generates outputs close to the observed values, while taking into account possible variation due to different individuals and possible heterogeneity between the Fibrosis severity groups. Moreover, as the original measurements are concentrations, in order to normalize and stabilize variance, we apply a log transformation to the output of the model and use the observations in the log scale.

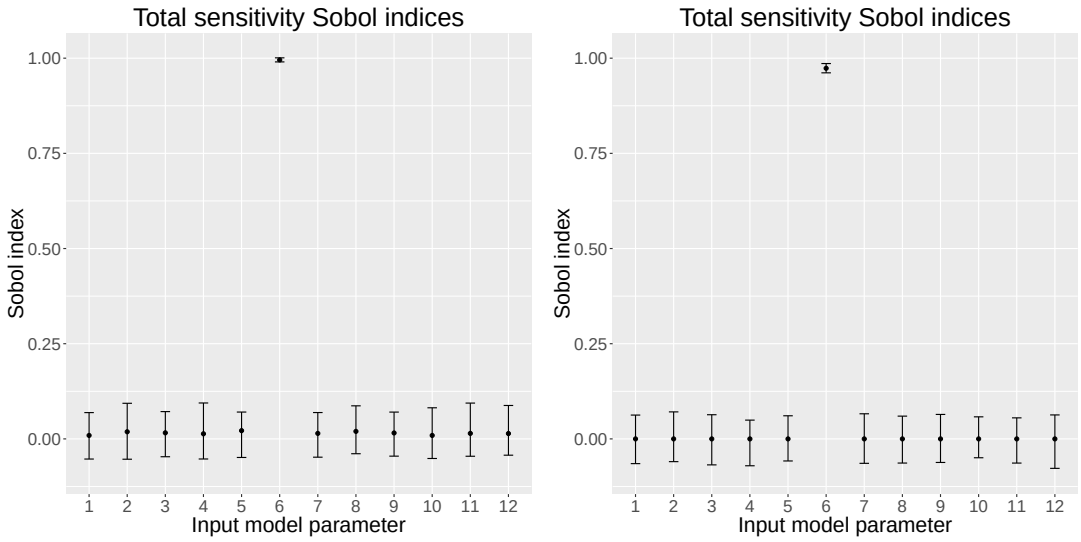
Based on this, to specify the maximum likelihood criterion. We assume parametric model

$$\eta_{i,j} = M_j(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}}) + g_j + \epsilon_{i,j};\ i = 1, ..., m_j, j = 1, ..., 3 \tag{10}$$

$$g_j \sim \mathrm{N}(0, \tau_j); \tag{11}$$

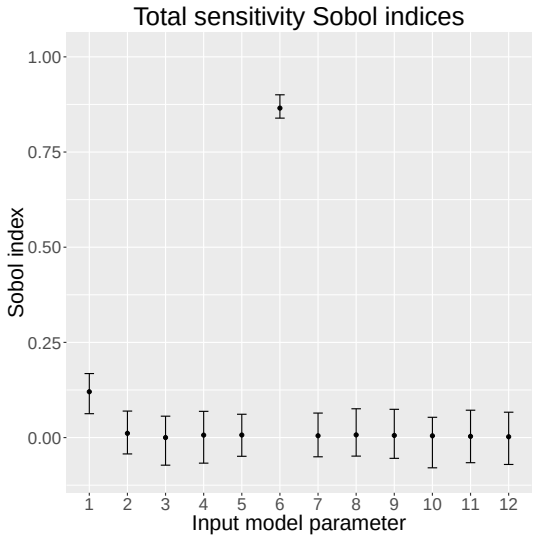
$$\epsilon_{i,j} \sim \mathrm{N}(0, \tau),$$

where $M_j(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}})$ is the uMCP1 real output (in log scale) of the simulator at the λ_{ET} and λ_{PM} input parameter values, and $\eta_{i,j}$ is the uMCP1 ($ng/mgcr$) measurement (in log scale), of the i-th individual of the j-th severity level. The fibrosis severity levels were labeled as $j = 1$ for low level, $j = 2$ for intermediate level, and $j = 3$ for high level. The size of each fibrosis sensitivity group was $m_1 = 23$, $m_2 = 33$, and $m_3 = 28$. Notice that the model (10) allows for different variations among different severity levels. In particular, the error term $\epsilon_{i,j}$ accounts for the noise variation, while g_j accounts for the different variability accross different severity levels. Precisely, the resulting likelihood function of vector $\eta = (\eta_{i,j} | i = 1, ..., m_j; j = 1, ..., 3)$



(a) Low fibrosis sensitivity level

(b) Intermediate fibrosis sensitivity level



(c) High fibrosis sensitivity level

Figure 3: Total sensitivity Sobol indices estimates and confidence intervals. The values on the horizontal axis indicate the indices of vector $(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{TM}}, \lambda_{\mathrm{GM}}, \lambda_{\mathrm{QM}}, \lambda_{\mathrm{QrM}}, \lambda_{\mathrm{PM}}, \lambda_{\rho\mathrm{f}}, \lambda_{\rho\mathrm{m}}, \lambda_{\rho\mathrm{T}}, \lambda_{\mathrm{fE}}, \lambda_{\mathrm{mfT}}, \lambda_{\mathrm{mfG}})$.

corresponds to a multivariate normal distribution density with mean such as $\mathrm{E}(\eta_{i,j}) = M_j(\lambda_{\mathrm{PM}}, \lambda_{\mathrm{ET}})$ and covariance matrix such as $\mathrm{Cov}(\eta_{i,j}, \eta_{i',j'}) = (\tau_j + \tau\delta_{i,i'})\delta_{j,j'}$, for $i = 1, ..., m_j$, $j = 1, 2, 3$.

The calibration is performed, by maximizing the likelihood function that leads to the optimization problem

$$(\lambda^*_{\mathrm{ET}}, \lambda^*_{\mathrm{PM}}, \tau^*_0, \tau^*_1, \tau^*_2, \tau^*) = \arg \min_{\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}}, \tau_0, \tau_1, \tau_2, \tau} f(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}}, \tau_0, \tau_1, \tau_2, \tau) \tag{12}$$

where

$$\begin{aligned} f(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}}, \tau_0, \tau_1, \tau_2, \tau) = \frac{1}{2}\sum_{j=0}^{2} \log\det(\Sigma_j(\tau_j, \tau)) \\ + \frac{1}{2}\sum_{j=0}^{2} (\eta_j - M_j(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}}))^\top [\Sigma_j(\tau_j, \tau)]^{-1} (\eta_j - M_j(\lambda_{\mathrm{ET}}, \lambda_{\mathrm{PM}})), \end{aligned} \tag{13}$$

is the objective function with $\Sigma_j = \tau_j 1_{n_j} + \tau \mathbb{I}_{n_j}$, for $j = 0, 1, 2$.

The optimization problem (12) is solved via Bayesian global optimization (BGO), because the objective function (13) is prohibitively expensive to be computed massively as it requires one to run the expensive simulator. Hence, according to BGO, a GP emulator is used as a proxy of (13). For the GP emulator, the prior mean was specified to be constant $h(x) = 1$, the covariance function was specified to be a tensored 5/2-Matern kernel, and the nugget term was specified to be $\kappa = 0.001$ after pilot runs to check the stability of the computations. The acquisition function considered was the Expected Improvement. We considered an initial training data-set of size 15, whose input values were chosen via LHS. The BGO algorithm ran for 51 iterations. The EI was optimized by PISAA [9] optimizer. In particular, EI was evaluated at 100 random locations and the 'best' values were used as seeds for the PISAA optimizer which performed 1000 iterations targeting IE. At each iteration, we increase the training data-set by adding 10 new objective function evaluations at the locations suggested by the highest local maxima of the Expected Improvement accordingly. At the end the procedure was trained against a data-set of 525 simulations.

In Figure 4, we present convergence plots of BGO implemented for the solution of (12). The evolution of the best value $f(x^*)$ discovered is presented in Figure 4a. The optimal values of the calibration parameters were detected to be $\lambda^*_{\mathrm{ET}} = 2.3759$ and $\lambda^*_{\mathrm{PM}} = 0.00127$, while the estimated nuisance parameters where $\tau^* = 1.0192$, $\tau^*_0 = 0.052$, $\tau^*_1 = 0.605$, $\tau^*_2 = 0.83$. The best objective function value discovered was $f(x^*) = 48.01$, at $x^* = (\lambda^*_{\mathrm{ET}}, \lambda^*_{\mathrm{PM}}, \tau^*_0, \tau^*_1, \tau^*_2, \tau^*)$. These parameters are in the reasonable ranges listed in [5] but with a better fitted value, $f(x^*) = 48.01$. Moreover, we observe that our analysis suggests that the variation of MCP-1 concentration may increase with the level of severity since the estimates of the nuisance parameters $\{\tau^*_j\}_{j=0}^2$ increase with j.

The predicted standard deviation of the new location against the iteration of the algorithms is presented in Figure 4b, and we observe that it is acceptably small when it terminates.

We performed an informal accuracy assessment of the derived estimates of the tunable parameters λ^*_{ET} and λ^*_{PM}. The square root mean square error of the output of the fibrosis simulator using the calibrated values $\lambda^*_{\mathrm{ET}} = 2.3759$ and $\lambda^*_{\mathrm{PM}} = 0.00127$; $\mathrm{smse}(\lambda^*_{\mathrm{ET}}, \lambda^*_{\mathrm{PM}}) = \sqrt{\frac{1}{84}\sum_{i,j}(\eta_{i,j} - M_j(\lambda^*_{\mathrm{ET}}, \lambda^*_{\mathrm{PM}}))^2}$ is 0.1157. In Figure 5, we present the distribution of the squared root predictive mean squared errors produced by a Monte Carlo cross-validation where the observed data-set was randomly splinted 30 times into validation data-sets that consist of 15 observations (5 from each level) and training data-set that consist of the rest

observations.

Here, we used this approach instead of K&O[10] due to its efficient mechanism that adaptively selects training-data from areas that both reduce uncertainty and optimize the simulator. This allows the algorithm to efficiently collect more desirable training data for a given computational budget. K&O is more appropriate for cases where the main interest lies on the design of an emulator, as well as it may present non-identifiability between the discrepancy term and calibration parameters which is difficult to be addressed. We could also account for discrepancy if g_j was parametrised with a non-zero mean in (11), however here the discrepancy is considered as negligible.

Discussion In our sensitivity analysis and calibration studies, we show that the parameter λ_{PM} is the most sensitive parameter for three different patient groups (low, intermediate and high fibrosis). This is consistent with the treatment studies in [5]: the treatment with anti-MCP-1, decreasing MCP-1 concentration, is a beneficial therapeutic plan for renal fibrosis [25]. The second most sensitive parameter is λ_{ET} which is associated with the production of TEC by TGF-β and plays an important role in renal fibrosis. This is validated by the anti-TGF-β treatment, a specific and effective therapy for chornic kidney disease associated with renal fibrosis [13].

The calibration approach is based on the available patient data which has three classification groups by fibrosis severity, “low", “intermediate" and “high." The calibration result provides a guidance to weight the network shown in 1 : MCP-1 plays the most important role in the renal fibrosis due to the inflammation; the TEC production by TGF-β is also crucial in the fibrosis progression. Therefore, once more clinical/biological data becomes available, the interaction mechanism between TEC, MCP-1 and TGF-β needs to be explored more carefully

5 Conclusion

In this paper, we have proposed a general calibration approach for biological systems. This approach is based on surrogate modeling, global sensitivity analysis, and the Gaussian process. It has been applied to a mathematical model of renal tubulointerstitial fibrosis developed in Hao et al. [5]. In our analysis we found that the parameter λ_{PM} which is associated with anti-MCP-1 is the most sensitive parameter for three different patient groups (low, intermediate and high fibrosis), while λ_{ET} which is associated with the production of TEC by TGF-β is the second most sensitive parameter. Although these computational results do not provide more information in modeling point of view, they guide us to explore the “high" sensitivity group once more clinical/biological data becomes available which gives us a direction on modeling refinement. Moreover, this general approach can be generalized to other biological models and provides a systematical way to quantify the precision medicine and personalized treatments guided by computational modeling.

Acknowledgments

Hao’s research was supported by the American Heart Association (Grant 17SDG33660722) and the National Science Foundation (Grant DMS-1818769). Lin gratefully acknowledge the support from the National Science Foundation (DMS-1555072, DMS-1736364, and DMS-1821233).

References

[1] Milton Abramowitz, Irene A Stegun, et al. Handbook of mathematical functions. *Applied mathematics series*, 55:62, 1966.

[2] GEB Archer, A Saltelli, and IM Sobol. Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58(2):99–120, 1997.

[3] Robert B Gramacy and Herbert KH Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.

[4] W. Hao, C. Marsh, and A. Friedman. A Mathematical Model of Idiopathic Pulmonary Fibrosis. *PLoS ONE*, 10(9):e0135097, 2015.

[5] W. Hao, B. H. Rovin, and A. Friedman. Mathematical model of renal interstitial fibrosis. *Proc. Natl. Acad. Sci. U.S.A.*, 111(39):14193–14198, Sep 2014.

[6] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325, 1948.

[7] Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. Uncertainties assessment in global sensitivity indices estimation from metamodels. *International Journal for Uncertainty Quantification*, 4(1), 2014.

[8] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[9] Georgios Karagiannis, Bledar A. Konomi, Guang Lin, and Faming Liang. Parallel and interacting stochastic approximation annealing algorithms for global optimisation. *Statistics and Computing*, pages 1–19, 2016.

[10] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[11] Loic Le Gratiet, Claire Cannamela, and Bertrand Iooss. A bayesian approach for global sensitivity analysis of (multifidelity) computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):336–363, 2014.

[12] W. J. Conover M. D. McKay, R. J. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[13] X. M. Meng, P. M. Tang, J. Li, and H. Y. Lan. Tgf-β/smad signaling in renal fibrosis. *Frontiers in physiology*, 6:82, 2015.

[14] Hervé Monod, Cédric Naud, and David Makowski. Uncertainty and sensitivity analysis for crop models. *Working with dynamic crop models: Evaluation, analysis, parameterization, and applications*, 4:55–100, 2006.

[15] Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[16] Jeremy E Oakley and Anthony O'Hagan. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

[17] Anthony O'Hagan and JFC Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–42, 1978.

[18] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[19] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[20] Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 877–885, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[21] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280, 2001.

[22] I.M. Sobol. Sensitivity estimates for non linear mathematical models. *Math. Modeling Comput. Experiment*, 1:407–414, 1993.

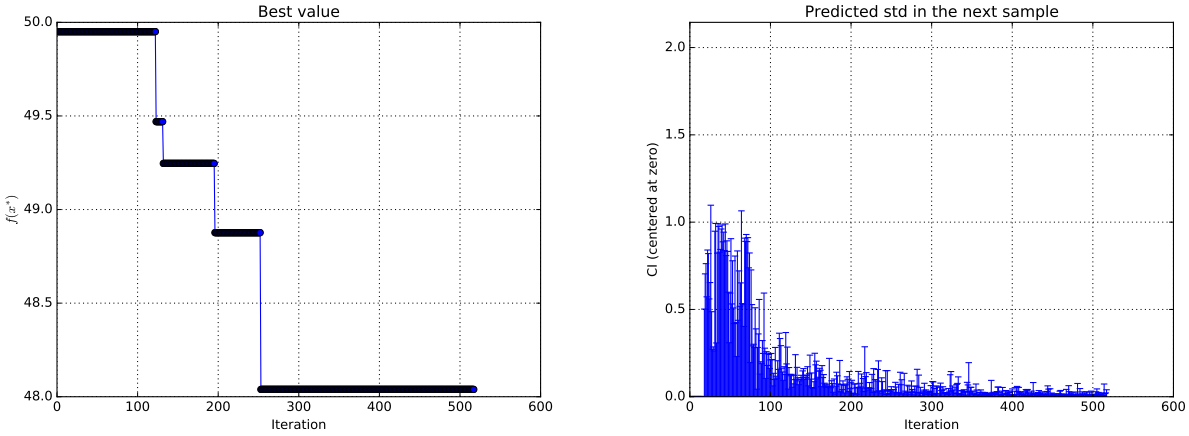
[23] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[24] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with student-t likelihood. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1910–1918. Curran Associates, Inc., 2009.

[25] T. Wada, K. Furuichi, N. Sakai, Y. Iwata, K. Kitagawa, Y. Ishida, T. Kondo, H. Hashimoto, Y. Ishiwata, N. Mukaida, N. Tomosugi, K. Matsushima, K. Egashira, and H. Yokoyama. Gene therapy via blockade of monocyte chemoattractant protein-1 for renal fibrosis. *J. Am. Soc. Nephrol.*, 15(4):940–948, Apr 2004.

[26] Xiaoliang Wan and George Em Karniadakis. Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM Journal on Scientific Computing*, 28(3):901–928, 2006.

[27] Xiu Yang, Minseok Choi, Guang Lin, and George Em Karniadakis. Adaptive anova decomposition of stochastic incompressible and compressible flows. *Journal of Computational Physics*, 231(4):1587–1614, 2012.



(a) Trace plot of the best values

(b) Trace plot

Figure 4: Convergence plots of the Bayesian global optimization

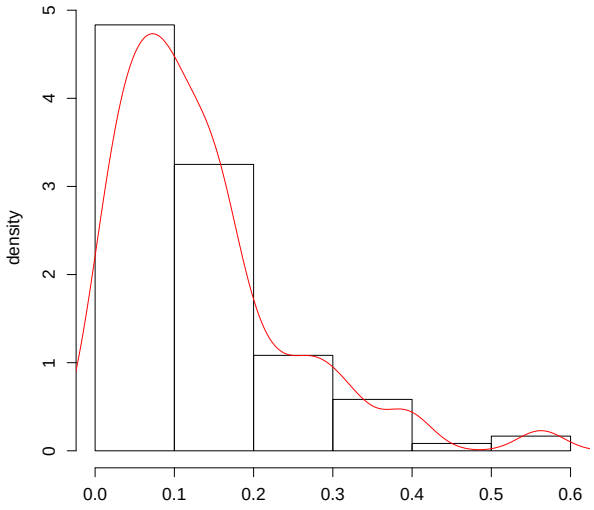


Figure 5: Squared root predictive mean squared errors by Monte Carlo cross validation with 30 realizations spliting the data-set into validation data-set that consists of 15 observations and training data-set of 69 observations.