Introduction to Computer Experiments

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA
Version: August 16, 2018, 13:13

1 Introduction

It is now common in science (physics, astronomy, geology, chemistry, *etc.*) to study physical systems by constructing complex computer-resident mathematical models intended to simulate those systems, and to try to learn about the physical system or to predict its later behavior by exploring how the computer simulation model behaves.

For now let's consider only the simulation of non-random scalar (univariate) quantities $Y(x)$, that may depend on a vector $x \in \mathcal{X} \subset \mathbb{R}^d_{\text{M}}$ of d observable (and perhaps even adjustable) quantities such as location, time, pressure, temperature, *etc.* Typically measurements of the physical system will entail measurement error of some kind, so we can write the field observations as:

$$Y_i^{\text{F}} = Y(x_i) + \epsilon_i, \qquad i \in I^{\text{F}}$$

Most commonly one takes the $\{\epsilon_i\}$ to be independent and normally distributed, but possibly heteroskedastic. No serious complications arise in allowing them to have an arbitrary but known covariance Σ^ϵ, or even a known correlation R^{F}_ϵ but uncertain precision λ_ϵ, so $\Sigma^\epsilon = \lambda_\epsilon^{-1} R^{\text{F}}_\epsilon$.

Without a mathematical model there is no way to predict what $Y(x)$ might be at some untried point $x \in \mathcal{X}$. A computer simulation model, or "Simulator", is a computer program intended to generate predictions $Y^{\text{M}}(x)$ of $Y(x)$, usually based on an underlying mathematical model (for example, the computer code might generate numerical solutions to a partial differential equation (PDE) whose initial conditions, boundary conditions, and perhaps other features are somehow determined by x). Most often Simulators require the specification of additional input quantities $t \in \mathcal{T}$ that might include "tuning" parameters or physical constants that might be variable in the Simulator but not in nature. Thus the model output will be $Y^{\text{M}}(x, t)$, which is intended to be close to $Y(x)$.

But:

- How close is it, at the input locations $\{x_i\}$ where we have data? *and,*
- How close can we expect it to be, at an untried point x? *and,*
- At what new input points $\{x_j\}$ should we run the Simulator to learn more about it most efficiently?

These are some of the questions that led to the study of the Calibration and Validation of Computer Models.

2 Gaussian Processes

Without some kind of smoothness assumptions the problem is hopeless. If the dependence of values of the physical system $\{Y(x_i)\}$ on their inputs $\{x_i\}$ is wild or chaotic, then observations at some locations won't help guide predictions at others, and we can learn nothing from the successes and failures of a Simulator at some locations about its performance at others. Often, however, continuity is a natural feature to expect— that near-by points $x_1, x_2 \in \mathcal{X}$ will lead to similar values $Y(x_1), Y(x_2)$ of the system and to similar Simulator outputs $Y^{\mathrm{M}}(x_1, t_1), Y^{\mathrm{M}}(x_2, t_2)$ for similar t_1, t_2. Beginning in the 1980s and 1990s a number of authors (e.g., Sacks et al., 1989; Currin et al., 1991; Kennedy and O'Hagan, 2000, 2001; Higdon et al., 2004; Heitmann et al., 2006; Bayarri et al., 2007b; Higdon et al., 2008; Bayarri et al., 2009, 2015, *etc.*) have pursued the theme of treating the values of $Y(x)$ (and even of $Y^{\mathrm{M}}(x,t)$) at untried points $x \in \mathcal{X}$ (and of $t \in \mathcal{T}$) as a collection of *random variables* indexed by $x \in \mathcal{X}$ and $t \in \mathcal{T}$, whose joint distribution is partially unknown and about which we might learn from data.

The most common approach is to construct *two* random field models: one "$Y^{\mathrm{M}}(x,t)$", the *Emulator*, intended to replicate the Simulator model however well or badly it succeeds in simulating nature; and another "$\delta_t(x)$", the *Discrepancy*, intended to model the difference between the Simulator and the natural system $Y(x)$. Gaussian Processes (GPs) are used for each of these, but with some differences. Typically the mean functions for both the Emulator and Discrepancy are taken to be linear functions of known basis functions, $\mu_{\mathrm{M}}(x,t) = \sum_{k \in K_{\mathrm{M}}} \psi_k(x,t)\beta_k^{\mathrm{M}} = \Psi_{\mathrm{M}}(x,t)\beta^{\mathrm{M}}$ and $\mu^\delta(x) = \sum_{k \in K^\delta} \psi_k(x)\beta_k^\delta = \Psi^\delta(x)\beta^\delta$, intended to pick up any systematic trends that would otherwise conflict with the isotropy of Y^{M} and the discrepancy between Y^{M} and Y, respectively. In the absence of concern about such trends or prior experience about discrepancies, it's common to take a single term $K_{\mathrm{M}} = \{0\}$ with $\psi_0 \equiv 1$ for a constant mean $\mu_{\mathrm{M}}(x,t) \equiv \beta_0$ for the emulator, and empty $K^\delta = \emptyset$ for zero discrepancy mean $\mu^\delta(x) \equiv 0$, but more elaborate choices don't add much more difficulty.

3 Emulation In Three Stages

3.1 Univariate Model Without Discrepancy

First we consider the problem of making predictions and inference upon observing the real-valued outcomes $y_i = Y^{\mathrm{M}}(x_i)$ of a computer simulation at a collection of *design points* $\mathcal{D} = \{x_i : \ i \in I^D\} \subset \mathcal{X}$ (more succinctly, upon observing $\mathbf{y} = Y^{\mathrm{M}}(\mathcal{D})$). Typically the design points are taken from a space-filling maxi-min Latin Hypercube design (LHD) (see Tang, 1993). Often the objective is to make predictions of simulation outcomes $\{Y^{\mathrm{M}}(x_i) : \ i \in I^S\}$ at a set $\mathcal{S} = \{x_i : \ i \in I^S\} \subset \mathcal{X}$ of untried input vectors, along with estimates of the prediction precision. All input components are usually transformed and scaled to the unit interval, making $\mathcal{X} = [0,1]^d$ a hypercube.

We model Y^{M} as a GP with mean of regression form $\mu_{\mathrm{M}}(\mathcal{D}) = \mathsf{E}Y^{\mathrm{M}}(\mathcal{D}) = \Psi_{\mathrm{M}}(\mathcal{D})\beta^{\mathrm{M}} = \sum_{k \in K} \psi_k^{\mathrm{M}}(x_i)\beta_k^{\mathrm{M}}$ for a vector $\Psi_{\mathrm{M}}(x)$ of $p = \#(K)$ specified basis functions $\psi_k^{\mathrm{M}}(x)$ and uncertain regression coefficients β_k^{M}, and with stationary covariance of the form $\lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}(x, x')$ with unknown precision λ_{M} and with stationary correlation matrix $R_{\mathrm{M}}(x, x') = r_{\mathrm{M}}(x - x' \mid \theta)$ from a specified parametric family governed by an unknown parameter vector $\theta \in \Theta$. The most common choice for correlation function is the separable power exponential $r_{\mathrm{M}}(h) := \exp\{-\sum |h_j/\ell_j|^\alpha\}$ with uncertain length scales $\{\ell_j\}$. Because the data are seldom informative about the power $\alpha \in [1,2]$, this smoothness parameter is usually specified to be $\alpha = 2$ or, for numerical reasons, something just a bit smaller like $\alpha = 1.9$, leaving $\theta = \{\ell_j\}$. It is possible, but seldom useful, to allow α_j to vary across dimensions. An appealing alternative to the power exponential is the product of Matérn covariances with smoothness parameter fixed at 3/2 or 5/2, with covariance functions:

$$r_{\mathrm{M}}^{3/2}(h \mid \theta) = e^{-\sum_{j\in J}\frac{h_j}{\ell_j}} \prod_{j\in J}\Big[1+\frac{h_j}{\ell_j}\Big] \qquad r_{\mathrm{M}}^{5/2}(h \mid \theta) = e^{-\sum_{j\in J}\frac{h_j}{\ell_j}} \prod_{j\in J}\Big[1+\frac{h_j}{\ell_j}+\Big(\frac{h_j}{\ell_j}\Big)^2/3\Big]. \tag{1}$$

Although other options are possible, it is customary to accord β^{M} and λ_{M} independent conjugate prior distributions $\beta^{\mathrm{M}} \sim \mathsf{No}(m_\beta, \lambda_\beta^{-1} I)$ and $\lambda_{\mathrm{M}} \sim \mathsf{Ga}(a_{\mathrm{M}}, b_{\mathrm{M}})$ (where $\mathsf{Ga}(\alpha,\beta)$ denotes the gamma distribution with *rate* β, or mean α/β), often with $\lambda_\beta = 0$ and $a_{\mathrm{M}} = b_{\mathrm{M}} = 0$ for the *reference* or weakly-informative prior distributions $\beta^{\mathrm{M}} \sim d\beta^{\mathrm{M}}$ and $\lambda_{\mathrm{M}} \sim \lambda_{\mathrm{M}}^{-1}\, d\lambda_{\mathrm{M}}$. Sometimes reference distributions suffice for θ as well. The model then is:

$$\begin{aligned}
Y^{\mathrm{M}}(\cdot) &\sim \mathsf{No}\big(\mu_{\mathrm{M}},\ \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}\big) && \text{(stationary GP)}\\
\mu_{\mathrm{M}}(x) &= \Psi_{\mathrm{M}}(x)\beta^{\mathrm{M}} && \text{(linear regression)}\\
R_{\mathrm{M}}(x,x') &= r_{\mathrm{M}}(x-x' \mid \theta) &&\\
r_{\mathrm{M}}(h \mid \theta) &= \exp\Big\{-\sum_{j\in J} |h_j/\ell_j|^\alpha\Big\} && \text{(or, alternately, Matérn— see (1) above)}\\
\beta^{\mathrm{M}} &\sim d\beta^{\mathrm{M}} && \text{(improper conjugate reference)}\\
\lambda_{\mathrm{M}} &\sim \lambda_{\mathrm{M}}^{-1} d\lambda_{\mathrm{M}} && \text{(improper conjugate reference)}\\
\theta &\sim \pi(d\theta) && \text{(arbitrary, maybe with density } \pi(\theta))
\end{aligned}$$

Upon observing the model output $\mathbf{y}^{\mathrm{M}} = Y^{\mathrm{M}}(\mathcal{D})$ at the n design points, each d-dimensional, the *conditional* mean and variance at a new input point x are

$$\hat{y}(x \mid \mathbf{y}^{\mathrm{M}}, \beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta) = \mu_{\mathrm{M}}(x) + R_{\mathrm{M}}(x,\mathcal{D}) R_{\mathrm{M}}^{-1}(\mathcal{D})\big[\mathbf{y}^{\mathrm{M}} - \mu_{\mathrm{M}}(\mathcal{D})\big] \tag{2a}$$

$$\hat{\sigma}^2(x \mid \mathbf{y}^{\mathrm{M}}, \beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta) = \lambda_{\mathrm{M}}^{-1}\Big[1 - R_{\mathrm{M}}(x,\mathcal{D}) R_{\mathrm{M}}^{-1}(\mathcal{D}) R_{\mathrm{M}}(\mathcal{D},x)\Big] \tag{2b}$$

and the predictive distribution is $Y^{\mathrm{M}}(x) \mid \mathbf{y}^{\mathrm{M}}, \beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta \sim \mathsf{No}(\hat{y}, \hat{\sigma}^2)$, where $\mu_{\mathrm{M}}(x) = \Psi_{\mathrm{M}}(x)\beta^{\mathrm{M}}$ and $\mu_{\mathrm{M}}(\mathcal{D}) = \Psi_{\mathrm{M}}(\mathcal{D})\beta^{\mathrm{M}}$ are the unconditional (or prior) means and where

$$R_{\mathrm{M}}(\mathcal{D}) = \begin{bmatrix} 1 & r_{\mathrm{M}}(x_1 - x_2 \mid \theta) & \cdots & r_{\mathrm{M}}(x_1 - x_n)\\ r_{\mathrm{M}}(x_2 - x_1 \mid \theta) & 1 & \cdots & r_{\mathrm{M}}(x_2 - x_n)\\ \vdots & \cdots & r_{\mathrm{M}}(x_i - x_j) & r_{\mathrm{M}}(x_i - x_n)\\ r_{\mathrm{M}}(x_n - x_1) & \cdots & \cdots & 1 \end{bmatrix} \tag{3a}$$

$$R_{\mathrm{M}}(x,\mathcal{D}) = R_{\mathrm{M}}(\mathcal{D},x)' = \big[r_{\mathrm{M}}(x-x_1),\ \ r_{\mathrm{M}}(x-x_2),\ \ \cdots\ \ r_{\mathrm{M}}(x-x_n)\big] \tag{3b}$$

where $n := \#(\mathcal{D})$ denotes the number of design points. To improve numerical stability it is common to replace the ones on the diagonal of R_{M} with $1+\eta$ for some small "nugget" $\eta > 0$. The negative log likelihood function is

$$\begin{aligned}\ell(\beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta \mid \mathbf{y}^{\mathrm{M}}) = \tfrac{1}{2}\big\{ &\log|R_{\mathrm{M}}| - n\log\lambda_{\mathrm{M}}\\ &+ \lambda_{\mathrm{M}}[\mathbf{y}^{\mathrm{M}} - \mu_{\mathrm{M}}(\mathcal{D})]' R_{\mathrm{M}}^{-1}(\mathcal{D})[\mathbf{y}^{\mathrm{M}} - \mu_{\mathrm{M}}(\mathcal{D})]\big\}\end{aligned} \tag{4a}$$

and the negative log prior density is

$$-\log \pi(\beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta) = \log \lambda_{\mathrm{M}} - \log \pi(\theta) \tag{4b}$$

where $|R_{\mathrm{M}}|$ denotes the determinant of the positive-definite matrix R_{M}.

One way to proceed is to draw an MCMC stream of $\{(\beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta)\}$ using Eqn (4) and, from this, use Eqn (2) to generate a stream of predictive quantiles or means. A less computationally intensive approach is to find MLE or MAP estimates $\{(\hat{\beta}^{\mathrm{M}}, \hat{\lambda}_{\mathrm{M}}, \hat{\theta})\}$ from Eqn (4) and use this and Eqn (2) to construct "plug-in" predictive estimates. While quicker, these will systematically under-represent predictive uncertainty.

With their conjugate distributions specified, either or both of the parameters β^{M} and λ_{M} may be integrated in closed form to reduce the dimensionality of the problem. The MLEs for the regression coefficient β^{M} and precision λ_{M} from Eqn (4) are

$$\hat{\beta}^{\mathrm{M}} = [\Psi' R_{\mathrm{M}}^{-1} \Psi]^{-1} \Psi' R_{\mathrm{M}}^{-1} \mathbf{y}^{\mathrm{M}} \tag{5a}$$

$$\hat{\lambda}_{\mathrm{M}} = \frac{n}{(\mathbf{y}^{\mathrm{M}} - \Psi \hat{\beta}^{\mathrm{M}})' R_{\mathrm{M}}^{-1} (\mathbf{y}^{\mathrm{M}} - \Psi \hat{\beta}^{\mathrm{M}})} \tag{5b}$$

and the negative log likelihood of Eqn (4a) can be rewritten

$$\begin{aligned} \ell(\beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta \mid \mathbf{y}^{\mathrm{M}}) = \tfrac{1}{2}\big\{ & \log |R_{\mathrm{M}}| - n \log \lambda_{\mathrm{M}} \\ & + \lambda_{\mathrm{M}} (\mathbf{y}^{\mathrm{M}} - \Psi \hat{\beta}^{\mathrm{M}})' R_{\mathrm{M}}^{-1} (\mathbf{y}^{\mathrm{M}} - \Psi \hat{\beta}^{\mathrm{M}}) \\ & + \lambda_{\mathrm{M}} (\beta^{\mathrm{M}} - \hat{\beta}^{\mathrm{M}})' [\Psi' R_{\mathrm{M}}^{-1} \Psi] (\beta^{\mathrm{M}} - \hat{\beta}^{\mathrm{M}}) \big\} \end{aligned} \tag{6a}$$

where we have simplified the notation by writing Ψ and R_{M} for $\Psi_{\mathrm{M}}(\mathcal{D})$ and $R_{\mathrm{M}}(\mathcal{D})$. Integrating $\exp(-\ell(\cdot))$ w.r.t. β^{M} (with an improper uniform prior distribution), using $|\Psi' R_{\mathrm{M}}^{-1} \Psi| = |\Psi \Psi'| / |R_{\mathrm{M}}|$, and writing $\hat{\mu}_{\mathrm{M}}$ for $\Psi_{\mathrm{M}}(\mathcal{D}) \hat{\beta}$, now leads to a marginal negative log likelihood

$$\ell(\lambda_{\mathrm{M}}, \theta \mid \mathbf{y}^{\mathrm{M}}) = -\tfrac{1}{2}(n-p) \log \lambda_{\mathrm{M}} + \frac{\lambda_{\mathrm{M}}}{2} [\mathbf{y}^{\mathrm{M}} - \hat{\mu}_{\mathrm{M}}]' R_{\mathrm{M}}^{-1}(\mathcal{D}) [\mathbf{y}^{\mathrm{M}} - \hat{\mu}_{\mathrm{M}}] \tag{6b}$$

where p denotes the regression dimension, *i.e.*, the length of β^{M} or the number of columns of Ψ_{M}. A further integral of $\exp(-\ell(\cdot))$ w.r.t λ_{M}, with conjugate reference prior[1] distribution $\lambda_{\mathrm{M}} \sim \lambda_{\mathrm{M}}^{-1}\, d\lambda_{\mathrm{M}}$, leads to the marginal negative log likelihood for θ:

$$\ell(\theta \mid \mathbf{y}^{\mathrm{M}}) = \tfrac{1}{2}(n-p) \log \left\{ \tfrac{1}{2} [\mathbf{y}^{\mathrm{M}} - \hat{\mu}_{\mathrm{M}}]' R_{\mathrm{M}}^{-1}(\mathcal{D}) [\mathbf{y}^{\mathrm{M}} - \hat{\mu}_{\mathrm{M}}] \right\} \tag{6c}$$

with the reference prior distributions for β^{M} and λ_{M}. From this we may locate the MLE $\hat{\theta} = \operatorname{argmin} \ell(\theta \mid \mathbf{y}^{\mathrm{M}}) = \operatorname{argmax} \hat{\lambda}_{\mathrm{M}}$ or the MAP or, if necessary, use $\ell(\theta \mid \mathbf{y}^{\mathrm{M}})$ in a Metropolis/Hastings approach to draw a sequence $\{\theta^{(t)}\} \subset \Theta$ from the posterior distribution. The predictive distribution of $Y^{\mathrm{M}}(x)$ or of its mean $\mu_{\mathrm{M}}(x)$ at any $x \in \mathcal{X}$ is available from this stream, by evaluating $R_{\mathrm{M}}^{-1}(\mathcal{D})$,

[1] Or, as an alternative, Bayarri et al. (2007b, *p.* 152) and Paulo (2005, §4.3, 4.5) recommend $a_{\mathrm{M}} = 1$ and $b_{\mathrm{M}} = 1/5\hat{\lambda}_{\mathrm{M}}$, and argue that results are insensitive to this choice.

$R_{\mathrm{M}}(x,\mathcal{D})$, $\hat{\mu}_{\mathrm{M}}(\mathcal{D})$, $\hat{\beta}^{\mathrm{M}}$, and $\hat{\lambda}_{\mathrm{M}}$ each time step (all of which depend on $\theta^{(t)}$), then successively drawing

$$\begin{aligned}
\lambda_{\mathrm{M}} &\sim \mathsf{Ga}\Big(\frac{n-p}{2},\ \frac{n}{2\hat{\lambda}_{\mathrm{M}}}\Big)\\
\beta^{\mathrm{M}} &\sim \mathsf{No}\Big(\hat{\beta}^{\mathrm{M}},\ [\lambda_{\mathrm{M}}\Psi_{\mathrm{M}}(\mathcal{D})'R_{\mathrm{M}}^{-1}(\mathcal{D})\Psi_{\mathrm{M}}(\mathcal{D})]^{-1}\Big)\\
\hat{y}(x) &= \Psi_{\mathrm{M}}(x)'\beta^{\mathrm{M}} + R_{\mathrm{M}}(x,\mathcal{D})R_{\mathrm{M}}^{-1}(\mathcal{D})\big[\mathbf{y}^{\mathrm{M}} - \Psi_{\mathrm{M}}(\mathcal{D})\beta\big]\\
\hat{\sigma}^2(x) &= \lambda_{\mathrm{M}}^{-1}\Big[1 - R_{\mathrm{M}}(x,\mathcal{D})R_{\mathrm{M}}^{-1}(\mathcal{D})R_{\mathrm{M}}(\mathcal{D},x)\Big]\\
Y^{\mathrm{M}}(x) &\sim \mathsf{No}\big(\hat{y}(x),\ \hat{\sigma}^2(x)\big)
\end{aligned}$$

or, equivalently,

$$Y^{\mathrm{M}}(x) \sim t\big(\hat{y}(x), V; n-p\big)$$

from the non-central Student t distribution, with

$$\begin{aligned}
V &= \frac{n}{(n-p)\ \hat{\lambda}_{\mathrm{M}}}\Big[1 - R_{\mathrm{M}}(x,\mathcal{D})R_{\mathrm{M}}^{-1}(\mathcal{D})R_{\mathrm{M}}(\mathcal{D},x)\Big]\\
&= \frac{(\mathbf{y}^{\mathrm{M}} - \Psi\hat{\beta}^{\mathrm{M}})'R_{\mathrm{M}}^{-1}(\mathcal{D})(\mathbf{y}^{\mathrm{M}} - \Psi\hat{\beta}^{\mathrm{M}})}{n-p}\Big[1 - R_{\mathrm{M}}(x,\mathcal{D})R_{\mathrm{M}}^{-1}(\mathcal{D})R_{\mathrm{M}}(\mathcal{D},x)\Big].
\end{aligned}$$

3.2 Univariate Model With Discrepancy

When field observations Y^{F} are available of the system Y that Y^{M} was intended to simulate, a simple iid zero-mean measurement-error model for the difference $Y^{\mathrm{F}} - Y^{\mathrm{M}}$ is usually too simplistic to capture the difference between modeled and measured values. It is more realistic to expect that these "model discrepancies" will be correlated and that their means and variances will depend on $x \in \mathcal{X}$. One approach in the spirit of Section (3.1) is to model the difference between field observations $Y^{\mathrm{F}}(x)$ and model predictions $Y^{\mathrm{M}}(x)$ as a Gaussian Process $\delta(x)$ independent of $Y^{\mathrm{M}}(x)$, with its own mean and covariance function.

One new feature we face is that typically the computer model Y^{M} will have *more parameters* than the field observations do— technical things like the mesh size or time step used in a simulation, or unobserved quantities like a friction coefficient or a reaction rate that can be set to arbitrary values in a simulator but which have specific values in the field. The usual way to accommodate this is to expand the input parameter from a single vector x to a pair (x,t) for the computer models (both simulator and emulator) only: "x" for those parameter(s) that are observable and perhaps even adjustable in the field (like locations in space and time, or features of initial conditions) and "t" (mnemonic for "tuning parameter") for those parameter(s) that appear or vary only in the computer models. Sometimes to shorten formulas we will write $x^* \equiv (x,t)$, with components $x^*_j = x_j$ for $j \in J_x$ and $x^*_j = t_j$ for $j \in J_t$ the appropriate subsets of the index set $J = J_x \cup J_t$. Denote by "$t_\star$" a nominal value of the tuning parameter $t \in \mathcal{T}$, usually described as "best" vector in that $Y^{\mathrm{M}}(\mathcal{F}, t_\star)$ is nearest to $Y^F(\mathcal{F})$ in some sense, or sometimes (optimistically) described as the "true" value and, for any $x \in \mathcal{F}$, denote by $x^\star = (x, t_\star)$ the element $x^* \in \mathcal{X} \times \mathcal{T}$ formed by augmenting x with this nominal $t_\star$. From the equation "$Y(x) = Y^{\mathrm{M}}(x,t_\star) + \delta(x)$" below we see that the choice of the nominal tuning parameter $t_\star$ is completely confounded with that of the discrepancy function $\delta(x)$, *i.e.*, neither is statistically identifiable. We convey that by denoting the discrepancy "$\delta_{t_\star}(x)$". The intention is to make δ in some sense small by selecting a suitable $t_\star$. The dimensions of input and regression vectors are:

Field observations	$\{x_i\}$	f	$= \#(I^{\mathrm{F}}) = \#(\mathcal{F})$
Design points	$\{x_i\}$	n	$= \#(I^{D}) = \#(\mathcal{D})$
Regressors, Model	$\{\beta_k\}$	p_{M}	$= \#(K_{\mathrm{M}})$
Regressors, Field	$\{\beta_k\}$	p^{F}	$= \#(K^{\mathrm{F}})$
Regressors, Total	$\{\beta_k\}$	p	$= \#(K \equiv K_{\mathrm{M}} \cup K^{\mathrm{F}})$

The resulting hierarchical model then is:

$$
\begin{aligned}
\mathbf{y}_i^{\mathrm{F}} &= Y(x_i) + \epsilon_i && \text{(field observations with measurement error)}\\
Y(x_i) &= Y^{\mathrm{M}}(x_i, t_\star) + \delta_{t_\star}(x_i) && \text{(truth = model + discrepancy)}\\
Y^{\mathrm{M}}(x,t) &\sim \mathsf{No}\big(\mu_{\mathrm{M}},\ \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}\big) && \text{(stationary GPs)}\\
\delta_{t_\star}(x) &\sim \mathsf{No}\big(\mu^{\delta},\ \lambda_{\delta}^{-1} R_{\mathrm{M}}^{\delta}\big) && \\
\mu_{\mathrm{M}}(x,t) &= \Psi_{\mathrm{M}}(x,t)\beta^{\mathrm{M}} = \sum_{k\in K_{\mathrm{M}}} \psi_k(x,t)\beta_k && \text{(often } K_{\mathrm{M}} = \{0\},\ \psi_0 \equiv 1,\ \mu^{M} \equiv \beta_0^{\mathrm{M}}\text{)}\\
\mu^{\delta}(x) &= \Psi^{\delta}(x)\beta^{\delta} \qquad = \sum_{k\in K^{\mathrm{F}}} \psi_k(x)\beta_k && \text{(often } K^{\mathrm{F}} = \emptyset,\ \mu^{\delta} \equiv 0\text{)}\\
R_{\mathrm{M}}(x,t;x',t') &= r_{\mathrm{M}}(x-x', t-t' \mid \theta) && \text{(stationary correlations)}\\
R_{\mathrm{M}}^{\delta}(x,x') &= r^{\delta}(x-x' \mid \theta) && \\
r_{\mathrm{M}}(h \mid \theta) &= \exp\Big\{ -\sum_{j\in J} |h_j/\ell_j^{\mathrm{M}}|^{\alpha} \Big\} && \text{(power exponentials, or Matérn (1))}\\
r^{\delta}(h \mid \theta) &= \exp\Big\{ -\sum_{j\in J_x} |h_j/\ell_j^{\delta}|^{2} \Big\} && \text{(usually power = 2 for } \delta\text{)}\\
\epsilon_i &\sim \mathsf{No}(0, \Sigma_\epsilon) && \text{(take } \Sigma_\epsilon = \sigma_\epsilon^2 I \text{ for iid errors)}\\
\beta^{\mathrm{M}}, \beta^{\delta} &\sim d\beta^{\mathrm{M}}\ d\beta^{\delta} && \text{(improper conjugate uniform)}\\
\lambda_{\mathrm{M}} &\sim \mathsf{Ga}(a_{\mathrm{M}}, b_{\mathrm{M}}) && \text{(conjugate, \textit{e.g.}, } a_{\mathrm{M}} = 1,\ b_{\mathrm{M}} = 1/5\hat{\lambda}_{\mathrm{M}}\text{)}\\
\lambda^{\delta} &\sim \mathsf{Ga}(a^{\delta}, b^{\delta}) && \text{(conjugate, \textit{e.g.}, } a^{\delta} = 1,\ b^{\delta} = 1/5\hat{\lambda}^{\delta}\text{)}\\
\theta = \big\{\ell_j^{\mathrm{M}}, \ell_j^{\delta}\big\} &\sim \pi(d\theta) && \text{(arbitrary, maybe with density } \pi(\theta)\text{)}
\end{aligned}
$$

The two processes can be fit simultaneously, by first concatenating the input, output, and parameter vectors:

$$
\begin{aligned}
I &= I^{D} \cup I^{\mathrm{F}} \quad \text{(indexing design and field points } \mathcal{D} \cup \mathcal{F}\text{)}\\
J &= J_x \cup J_t \qquad \text{(indexing } d \text{ components of } x^* = (x,t)\text{)}\\
x^* &= (x', t')' = \big\{ x_j^* :\ j \in J \big\}\\
\beta &= [\beta^{\mathrm{M}}, \beta^{\delta}]\\
\Psi &= [\Psi_{\mathrm{M}}, \Psi^{\delta}]\\
\mu &= [\mu_{\mathrm{M}}, \mu^{\delta}] = \Psi\beta\\
\Sigma &= \begin{bmatrix} \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}(\mathcal{D}) & \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}(\mathcal{D}, \mathcal{F}) \\ \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}(\mathcal{F}, \mathcal{D}) & \lambda_{\mathrm{M}}^{-1} R_{\mathrm{M}}(\mathcal{F}) + \lambda_{\delta}^{-1} R_{\mathrm{M}}^{\delta}(\mathcal{F}) + \Sigma_\epsilon \end{bmatrix}\\
\mathbf{y} &= [\mathbf{y}^{\mathrm{M}}, \mathbf{y}^{\mathrm{F}}] \sim \mathsf{No}(\mu, \Sigma)
\end{aligned}
$$

where

$$\begin{aligned}
[R_{\mathrm{M}}]_{ij}(\mathcal{D}) &= r_{\mathrm{M}}(x_i^* - x_j^* \mid \theta),\ i \in I^D,\ \ j \in I^D \\
[R_{\mathrm{M}}]_{ji}(\mathcal{F}, \mathcal{D}) = [R_{\mathrm{M}}]_{ij}(\mathcal{D}, \mathcal{F}) &= r_{\mathrm{M}}(x_i^* - x_j^* \mid \theta),\ i \in I^D,\ \ j \in I^{\mathrm{F}} \\
]R_{\mathrm{M}}^{\delta}]_{ij}(\mathcal{F}) &= r^{\mathrm{F}}(x_i - x_j \mid \theta),\ \ i \in I^{\mathrm{F}},\ \ j \in I^{\mathrm{F}}.
\end{aligned}$$

The negative log likelihood is now that for a single $(n+f)$-dimensional multivariate normal model,

$$\ell(\beta, \lambda, \theta \mid \mathbf{y}^{\mathrm{M}},\, \mathbf{y}^{\mathrm{F}}) = \tfrac{1}{2}\big\{ \log|\Sigma| + [\mathbf{y} - \mu]'\Sigma^{-1}[\mathbf{y} - \mu] \big\}$$

and the negative log prior density is

$$\begin{aligned}
-\log \pi(\beta^{\mathrm{M}}, \lambda_{\mathrm{M}}, \theta) &= (1 - a_{\mathrm{M}}) \log \lambda_{\mathrm{M}} + b_{\mathrm{M}}\lambda_{\mathrm{M}} + (1 - a_\delta) \log \lambda_\delta + b_\delta \lambda_\delta - \log \pi(\theta) \\
&= (1/5\hat{\lambda}_{\mathrm{M}})\lambda_{\mathrm{M}} + (1/5\hat{\lambda}_\delta)\lambda_\delta - \log \pi(\theta),
\end{aligned}$$

using the suggested hyperparameter values. Bayarri et al. (2007b) recommend what they call "modularization", in which the design parameters β^{M} and ℓ^{M} are fit as in Section (3.1) using only the model output and not the field data, then with these parameters fixed the field data are used as above to generate posterior samples and predictive distributions. In this case the Sherman-Morrison-Woodbury formula may be used, substantially reducing the computational complexity of inverting Σ if (as usual) $f \ll n$.

3.3 Multivariate Model With Discrepancy

Theoretically nothing new is needed to extend Model Emulation to the multivariate setting in which Simulator model outputs consist of vectors in some Euclidean space $\mathbb{R}_{\mathrm{M}}^p$ or functions: simply let one input variable x_i be an index for which output variable or function argument is to be generated, or (better) let $(p-1)$ input variables specify a location on the simplex Δ_p to specify what affine linear combination of output dimensions should be returned. *In practice*, however, the similar approaches introduced independently by Higdon et al. (2008) and by Bayarri et al. (2007a) are far more successful: apply some dimension reduction technique to resolve a high-dimensional (say, $\mathbb{R}_{\mathrm{M}}^p$ valued) Simulator output into a modest-dimensional (say, r-dimensional with $r \ll p$) vector of (nearly) orthogonal responses, then construct r independent univariate emulators (which may all run in parallel) to generate the needed predictive means and variances. Higdon et al. used Principal Components Analysis (Wolpert, 2014; Mardia et al., 1979, Chap. 8) for orthogonal dimension reduction, while Bayarri et al. used discrete wavelet transforms with threshholding (see Vidakovic, 1999, §6.3, §8.3).

References

Bayarri, M. J., Berger, J. O., Cafeo, J. A., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007a), "Computer Model Validation with Functional Output," *Annals of Statistics*, 35, doi:10.1214/009053607000000163.

Bayarri, M. J., Berger, J. O., Calder, E. S., Dalbey, K., Lunagomez, S., Patra, A. K., Pitman, E. B., Spiller, E., and Wolpert, R. L. (2009), "Using Statistical and Computer Models to Quantify Volcanic Hazards," *Technometrics*, 51, 402–413, doi:10.1198/TECH.2009.08018.

Bayarri, M. J., Berger, J. O., Calder, E. S., Patra, A. K., Pitman, E. B., Spiller, E. T., and Wolpert, R. L. (2015), “Probabilistic Quantification of Hazards: A Methodology Using Small Ensembles of Physics Based Simulations and Statistical Surrogates,” 5, 297–339, doi:10.1615/Int.J.UncertaintyQuantification.2015011451.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J. C., Lin, C.-H., and Tu, J. (2007b), “A Framework for Validation of Computer Models,” *Technometrics*, 49, 138–154, doi:10.1198/004017007000000092.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), “Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments,” *Journal of the American Statistical Association*, 86, 953–963, doi:10.2307/2290511.

Heitmann, K., Higdon, D., Nakhleh, C., and Habib, S. (2006), “Cosmic Calibration,” *Astrophysical Journal Letters*, 646, L1–L4.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), “Computer Model Calibration using High-Dimensional Output,” *Journal of the American Statistical Association*, 103, 570–583, doi:10.1198/016214507000000888.

Higdon, D., Kennedy, M. C., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004), “Combining field observations and simulations for calibration and prediction,” *SIAM Journal on Scientific Computing*, 26, 448–466.

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, doi:10.1093/biomet/87.1.1.

Kennedy, M. C. and O’Hagan, A. (2001), “Bayesian calibration of computer models (with discussion),” *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 63, 425–464, doi:10.1111/1467-9868.00294.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, New York, NY: Academic Press.

Paulo, R. (2005), “Default priors for Gaussian processes,” *Annals of Statistics*, 33, 556–582, doi: DOI10.1214/009053604000001264.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), “Design and Analysis of Computer Experiments,” *Statistical Science*, 4, 409–435, doi:10.1214/ss/1177012413.

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

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, Computational & Graphical Statistics, New York, NY: John Wiley & Sons.

Wolpert, R. L. (2014), *Principal Components Analysis*, course lecture notes on PCA.