Bayesian Calibration of Inexact Computer Models

James Matuk

Research Group in Design of Physical and Computer Experiments

March 5, 2018

Overview

Calibration Assumptions

Let $\mathcal{X} \subset \mathbb{R}^d$ open and bounded.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \to \mathbb{R}$. There exist some $k \in \mathbb{N}$ so that $D^{(\alpha)}y(\cdot)$ exists and is bounded for all $\mathbb{R}^d$ vectors of non-negative integers α so that $\|\alpha\|_{L^1} \leq k$.

Calibration Assumptions

Let $\mathcal{X} \subset \mathbb{R}^d$ open and bounded.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \to \mathbb{R}$. There exist some $k \in \mathbb{N}$ so that $D^{(\alpha)}y(\cdot)$ exists and is bounded for all $\mathbb{R}^d$ vectors of non-negative integers α so that $\|\alpha\|_{L^1} \leq k$.

(ii) A computer model $f(\cdot, \cdot)$ is a deterministic map from $\mathcal{X} \times \mathbb{R} \to \mathbb{R}$ where $D^{(\alpha,0)}f(\cdot, \cdot)$ exists and is bounded.

Calibration Assumptions

Let $\mathcal{X} \subset \mathbb{R}^d$ open and bounded.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \to \mathbb{R}$. There exist some $k \in \mathbb{N}$ so that $D^{(\alpha)}y(\cdot)$ exists and is bounded for all $\mathbb{R}^d$ vectors of non-negative integers α so that $\|\alpha\|_{L^1} \leq k$.

(ii) A computer model $f(\cdot, \cdot)$ is a deterministic map from $\mathcal{X} \times \mathbb{R} \to \mathbb{R}$ where $D^{(\alpha,0)}f(\cdot, \cdot)$ exists and is bounded.

(iii) There exists a mapping L from the space of k differentiable functions defined on $\mathcal{X}$ to $\mathbb{R}$ so that there is some $\theta \in \Theta$ so that

$$L(y(\cdot) - f(\cdot, \theta)) < L(y(\cdot) - f(\cdot, t))$$

for any $t \in \Theta, t \neq \theta$

Model Bias

Define the model bias as

$$z_\theta(x) := y(x) - f(x, \theta).$$

Notice the bias is indexed by the 'true' or 'best' value of θ possible. So,

$$y(x) = f(x, \theta) + z_\theta(x).$$

Bayesian Model

Suppose we have observations $\mathbf{Y} = Y_1, \ldots, Y_n$ corresponding to inputs $\mathbf{x} = x_1, \ldots, x_n$ corrupted by some iid additive gaussian noise $\epsilon_1, \ldots, \epsilon_n$. i.e.

$$Y_i(x_i) = z(x_i) + \epsilon_i = f(x_i, \theta) + z_\theta(x_i) + \epsilon_i, \forall i = 1, \ldots, n.$$

Bayesian Model

Suppose we have observations $\mathbf{Y} = Y_1, \ldots, Y_n$ corresponding to inputs $\mathbf{x} = x_1, \ldots, x_n$ corrupted by some iid additive gaussian noise $\epsilon_1, \ldots, \epsilon_n$. i.e.

$$Y_i(x_i) = z(x_i) + \epsilon_i = f(x_i, \theta) + z_\theta(x_i) + \epsilon_i, \forall i = 1, \ldots, n.$$

So given,

$$\begin{aligned} Y_i | z_\theta(x_i), \theta &\overset{iid}{\sim} N(f(x_i, \theta) + z_\theta(x_i), v) \\ z_\theta(\cdot) | \theta &\sim GP(0, \sigma^2 r_\theta(x, x')) \\ \theta &\sim \pi(\theta) \end{aligned}$$

Bayesian Model

Suppose we have observations $\mathbf{Y} = Y_1, \ldots, Y_n$ corresponding to inputs $\mathbf{x} = x_1, \ldots, x_n$ corrupted by some iid additive gaussian noise $\epsilon_1, \ldots, \epsilon_n$. i.e.

$$Y_i(x_i) = z(x_i) + \epsilon_i = f(x_i, \theta) + z_\theta(x_i) + \epsilon_i, \forall i = 1, \ldots, n.$$

So given,

$$\begin{aligned} Y_i | z_\theta(x_i), \theta &\overset{iid}{\sim} N(f(x_i, \theta) + z_\theta(x_i), v) \\ z_\theta(\cdot)|\theta &\sim GP(0, \sigma^2 r_\theta(x, x')) \\ \theta &\sim \pi(\theta) \end{aligned}$$

we can find

$$\begin{aligned} \pi(\theta|Y) &\propto \int_{\mathbb{R}^n} \pi(\mathbf{Y}|z_\theta(\mathbf{x}))\pi(z_\theta(\mathbf{x})|\theta)\pi(\theta)d(z_\theta(\mathbf{x})) \\ \pi(z_\theta(x_0)|Y) &= \int_\Theta \pi(z_\theta(x_0)|\theta)\pi(\theta|\mathbf{Y})d\theta \end{aligned}$$

Loss Functions

Defining a loss function adds some additional structure to the problem.

(i) $L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\xi$

Loss Functions

Defining a loss function adds some additional structure to the problem.

(i) $L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\xi$

(ii) $L_{L^2(\mu)}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\mu(\xi)$

Loss Functions

Defining a loss function adds some additional structure to the problem.

(i) $L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\xi$

(ii) $L_{L^2(\mu)}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\mu(\xi)$

(iii) $L_{W_k^2}(y(\cdot) - f(\cdot, t)) = \sum_{\|\alpha\|_{L^1} \leq k} \|D^{(\alpha)} y(\cdot) - D^{(\alpha,0)} f(\cdot, t)\|_{L^2}^2$

Loss Functions

Defining a loss function adds some additional structure to the problem.

(i) $L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\xi$

(ii) $L_{L^2(\mu)}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\mu(\xi)$

(iii) $L_{W_k^2}(y(\cdot) - f(\cdot, t)) = \sum_{\|\alpha\|_{L^1} \le k} \|D^{(\alpha)} y(\cdot) - D^{(\alpha,0)} f(\cdot, t)\|_{L^2}^2$

(iv) $L_{W_k^2(\mu)}(y(\cdot) - f(\cdot, t)) = \sum_{\|\alpha\|_{L^1} \le k} \|D^{(\alpha)} y(\cdot) - D^{(\alpha,0)} f(\cdot, t)\|_{L^2(\mu)}^2$

The choice of loss (i) - (iv) will depend on the application and the information available.

L_{L^2} Example

The implications of assumptions (i) - (iii) and choice of loss function is that the bias should be orthogonal to the the the gradient of the computer model.

L_{L^2} Example

The implications of assumptions (i) - (iii) and choice of loss function is that the bias should be orthogonal to the the the gradient of the computer model. Consider,

$$L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (f(\xi, \theta) + z_\theta(\xi) - f(\xi, t))^2 d\xi.$$

L_{L^2} Example

The implications of assumptions (i) - (iii) and choice of loss function is that the bias should be orthogonal to the the the gradient of the computer model. Consider,

$$L_{L^2}(y(\cdot) - f(\cdot, t)) = \int_{\mathcal{X}} (f(\xi, \theta) + z_\theta(\xi) - f(\xi, t))^2 d\xi.$$

(Theorem 1 of [1]) Assuming all regularity conditions to exchange differentiaton and integration, then using standard optimality conditions one should enforce the following constraint

$$\int_{\mathcal{X}} D^{(0,1)} f(\xi, \theta) z_\theta(\xi) d\xi = 0.$$

General Orthogonality Condition

(Theorem 2 of [1]) For the most general loss considered,

$$L_{W_k^2}(y(\cdot) - f(\cdot, t)) = \sum_{\|\alpha\|_{L^1} \leq k} \|D^{(\alpha)} y(\cdot) - D^{(\alpha,0)} f(\cdot, t)\|_{L^2}^2,$$

under regularity conditions, one should enforce the following constraint,

$$\sum_{\|\alpha\|_{L^1} \leq k} \int_{\mathcal{X}} D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha,0)} z_\theta(\xi) d\mu(\xi) = 0.$$

General Orthogonality Condition

(Theorem 2 of [1]) For the most general loss considered,

$$L_{W_k^2}(y(\cdot) - f(\cdot, t)) = \sum_{\|\alpha\|_{L^1} \leq k} \|D^{(\alpha)} y(\cdot) - D^{(\alpha,0)} f(\cdot, t)\|_{L^2}^2,$$

under regularity conditions, one should enforce the following constraint,

$$\sum_{\|\alpha\|_{L^1} \leq k} \int_{\mathcal{X}} D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha,0)} z_\theta(\xi) d\mu(\xi) = 0.$$

These constraints can be enforced through the prior distribution on $z_\theta(\cdot)$.

Enforcing Orthogonality

Recall $z_\theta(\cdot)|\theta \sim GP(0, \sigma^2 r_\theta(x, x'))$

(Theorem 3 of [1]) If $r_\theta(x, x') = r(x, x') - h_\theta(x)^T H_\theta^{-1} h_\theta(x')$ with

Enforcing Orthogonality

Recall $z_\theta(\cdot)|\theta \sim GP(0, \sigma^2 r_\theta(x, x'))$

(Theorem 3 of [1]) If $r_\theta(x, x') = r(x, x') - h_\theta(x)^T H_\theta^{-1} h_\theta(x')$ with

$$
\begin{aligned}
h_\theta(x) &= \sum_{\|\alpha\|_{L^1} \leq k} \int_{\mathbb{R}^n} D^{(\alpha,1)} f(\xi, \theta) D^{(0,\alpha)} r(x, \xi) d\mu(\xi), \\
H_\theta &= \sum_{\|\alpha'\|_{L^1} \leq k} \sum_{\|\alpha\|_{L^1} \leq k} \int_{\mathcal{X}} \int_{\mathcal{X}} D^{(\alpha',1)} f(\xi, \theta) D^{(\alpha,1)} f(\xi, \theta) \\
&\times \quad D^{(\alpha',\alpha)} r(x, \xi) d\mu(\xi') d\mu(\xi),
\end{aligned}
$$

then with probability 1,

$$
\sum_{\|\alpha\|_{L^1} \leq k} \int_{\mathcal{X}} D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha,0)} z_\theta(\xi) d\mu(\xi) = 0.
$$

Enforcing Orthogonality

Recall $z_\theta(\cdot)|\theta \sim GP(0, \sigma^2 r_\theta(x, x'))$

(Theorem 3 of [1]) If $r_\theta(x, x') = r(x, x') - h_\theta(x)^T H_\theta^{-1} h_\theta(x')$ with

$$
\begin{aligned}
h_\theta(x) &= \sum_{\|\alpha\|_{L^1} \le k} \int_{\mathbb{R}^n} D^{(\alpha,1)} f(\xi, \theta) D^{(0,\alpha)} r(x, \xi) d\mu(\xi), \\
H_\theta &= \sum_{\|\alpha'\|_{L^1} \le k} \sum_{\|\alpha\|_{L^1} \le k} \int_{\mathcal{X}} \int_{\mathcal{X}} D^{(\alpha',1)} f(\xi, \theta) D^{(\alpha,1)} f(\xi, \theta) \\
&\times \quad D^{(\alpha',\alpha)} r(x, \xi) d\mu(\xi') d\mu(\xi),
\end{aligned}
$$

then with probability 1,

$$
\sum_{\|\alpha\|_{L^1} \le k} \int_{\mathcal{X}} D^{(\alpha,1)} f(\xi, \theta) D^{(\alpha,0)} z_\theta(\xi) d\mu(\xi) = 0.
$$

Notice that $r_\theta(x, x') = r(x, x') - h_\theta(x)^T H_\theta^{-1} h_\theta(x')$ takes a naive prior covariance function on the bias and updates it with gradient information from the computer model.

Enforcing Orthogonality Example

Suppose we have an input space of $x_1 = 1$, $x_2 = 2$ with $y(1) = 2.3, y(2) = 3.9$ and our biased model is given by

$$f(x,t) = t/4 + 2x + \sin(tx)$$

Enforcing Orthogonality Example

Suppose we have an input space of $x_1 = 1$, $x_2 = 2$ with $y(1) = 2.3, y(2) = 3.9$ and our biased model is given by

$$f(x, t) = t/4 + 2x + \sin(tx)$$

Under the Kennedy O'Hagan model a reasonable prior covarince conditional on θ is

$$cov_{KO}((z_\theta(1), z_\theta(2))^T | \theta) = \frac{1}{25} \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}$$

Enforcing Orthogonality Example

Suppose we have an input space of $x_1 = 1$, $x_2 = 2$ with $y(1) = 2.3, y(2) = 3.9$ and our biased model is given by

$$f(x,t) = t/4 + 2x + \sin(tx)$$

Under the Kennedy O'Hagan model a reasonable prior covarince conditional on θ is

$$cov_{KO}((z_\theta(1), z_\theta(2))^T|\theta) = \frac{1}{25}\begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}$$

The author assigns the reproducing Hilbert space norm as the loss function for this approach which is minimized at $\theta \approx -0.108$. This Loss function was not originally provided by Kennedy and O'Hagan, but attributed to them later.

Enforcing Orthogonality Example Continued

Using the this framework we work with L_{L^2} loss

$$(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2$$

which is minimized by $\theta \approx .022$.

Enforcing Orthogonality Example Continued

Using the this framework we work with L_{L^2} loss

$$(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2$$

which is minimized by $\theta \approx .022$. This implies

$$\frac{d}{dt}[(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2]|_{t=\theta=.022} = 0.$$

Enforcing Orthogonality Example Continued

Using the this framework we work with L_{L^2} loss

$$(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2$$

which is minimized by $\theta \approx .022$. This implies

$$\frac{d}{dt}[(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2]|_{t=\theta=.022} = 0.$$

further,

$$1.249 z_{\theta=.022}(1) + 2.248 z_{\theta=.022}(2) = 0.$$

Enforcing Orthogonality Example Continued

Using the this framework we work with L_{L^2} loss

$$(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2$$

which is minimized by $\theta \approx .022$. This implies

$$\frac{d}{dt}[(t/4 + 2 + \sin(t) - 2.3)^2 + (t/4 + 4 + \sin(2t) - 2.3)^2]|_{t=\theta=.022} = 0.$$

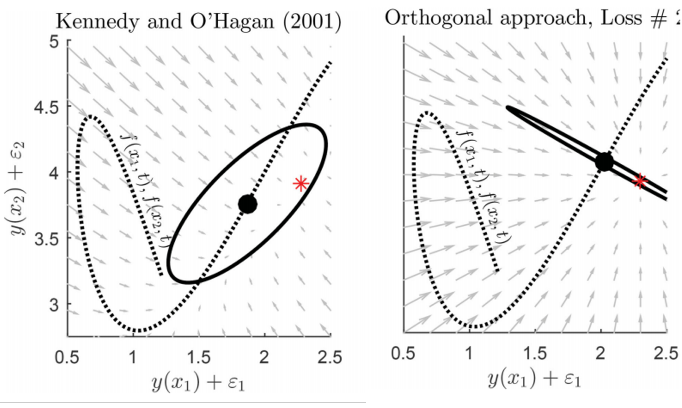
further,

$$1.249 z_{\theta=.022}(1) + 2.248 z_{\theta=.022}(2) = 0.$$

which is enforced through theorem 3 by making the prior covariance of the bias given θ,

$$cov_P((z_\theta(1), z_\theta(2))^T|\theta) \begin{bmatrix} 1.528 & -0.849 \\ -0.849 & 0.472 \end{bmatrix}$$

Enforcing Orthogonality Example Continued



The dot represents $(f(1, \theta), f(2, \theta))$, the ovals represent 95% credible regions for $(Y_1, Y_2)|\theta$, and the $*$ represents one draw from $(y(1) + \epsilon_1, y(2) + \epsilon_2)$.

Computing Difficult Integrals

Even for simpler loss functions like $L_{L^2(\mu)}$, integrals that define $r_\theta(x, x')$ are difficult to compute. However, one can draw a discrete set $(\xi_1, \ldots, \xi_N)$ independently from μ then use the following approximation,

$$L_{L^2(\mu)}(y(\cdot) - f(\cdot, t)) \approx \frac{1}{N} \sum_{i=1}^{N} (y(\xi_i) - f(\xi_i, t))^2$$

Computing Difficult Integrals

Even for simpler loss functions like $L_{L^2(\mu)}$, integrals that define $r_\theta(x, x')$ are difficult to compute. However, one can draw a discrete set $(\xi_1, \dots, \xi_N)$ independently from μ then use the following approximation,

$$L_{L^2(\mu)}(y(\cdot) - f(\cdot, t)) \approx \frac{1}{N}\sum_{i=1}^{N}(y(\xi_i) - f(\xi_i, t))^2$$

Let θ_N be a sequence of minimizers to the approximate loss, then $\theta_N \to \theta$ almost surely as $N \to \infty$. Using a plug-in estimator for θ his motivates setting

$$\begin{aligned} h_\theta(x) &= \frac{1}{N}\sum_{i=1}^{N} D^{(0,1)}f(\xi_i, \theta)r(x, \xi_i), \\ H_\theta &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} D^{(0,1)}f(\xi_i, \theta)D^{(0,1)}f(\xi, \theta)^T r(\xi_i, \xi_j). \end{aligned}$$

Model Emulation

Now suppose the computer model is computationally expensive so we wont have model evaluations or derivative information readily available. Assumptions (ii) and (iii) must be updated.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \rightarrow \mathbb{R}$.

Model Emulation

Now suppose the computer model is computationally expensive so we wont have model evaluations or derivative information readily available. Assumptions (ii) and (iii) must be updated.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \to \mathbb{R}$.

(ii) A computer model $f(\cdot, \cdot)$ follows a Gaussian process with mean $m_f(\cdot, \cdot)$ and covariance function $c_f(\cdot, \cdot)$. Then,

$$E_f[\int_{\mathcal{X}} (y(\xi) - f(\xi, t))^2 d\mu(\xi)] = \int_{\mathcal{X}} (y(\xi) - m_f(\xi, t))^2 + v_f(\xi, t) d\mu(\xi)$$

Model Emulation

Now suppose the computer model is computationally expensive so we wont have model evaluations or derivative information readily available. Assumptions (ii) and (iii) must be updated.

(i) A natural process $y(\cdot)$ is a deterministic map from $\mathcal{X} \to \mathbb{R}$.

(ii) A computer model $f(\cdot,\cdot)$ follows a Gaussian process with mean $m_f(\cdot,\cdot)$ and covariance function $c_f(\cdot,\cdot)$. Then,

$$E_f[\int_{\mathcal{X}} (y(\xi) - f(\xi,t))^2 d\mu(\xi)] = \int_{\mathcal{X}} (y(\xi) - m_f(\xi,t))^2 + v_f(\xi,t) d\mu(\xi)$$

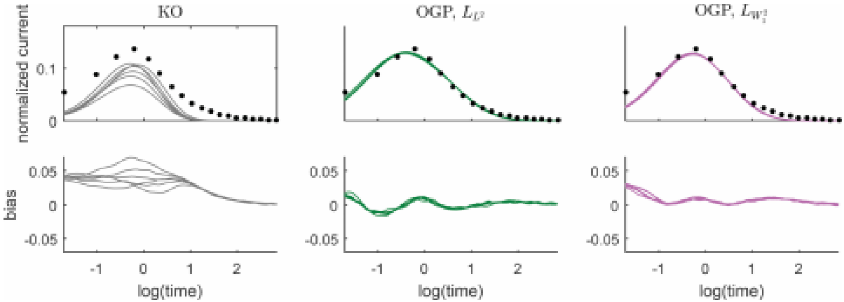
(iii) There exists some θ for which

$$\int_{\mathcal{X}} (y(\xi) - m_f(\xi,\theta))^2 + v_f(\xi,\theta) d\mu(\xi) < \int_{\mathcal{X}} (y(\xi) - m_f(\xi,t))^2 + v_f(\xi,t) d\mu(\xi)$$

for all $t \neq \theta$.

Ion Channel Example

The data set contains the current (response) needed for a sodium ion channel of a cardiac cell membrane to maintain a fixed amount (-35 mV) of membrane potential over time.



Summary

(i) The author formulates a method to specify the covariance structure of the model bias given θ using a loss function and optimality conditions.

Summary

(i) The author formulates a method to specify the covariance structure of the model bias given θ using a loss function and optimality conditions.

(ii) In some cases, this method seems to outperform the Kennedy, O'Hagan approach, but at a much greater computational cost. Particularly when the integrals $h_\theta(\cdot), H_\theta$ are not know in closed form.

References

Mathew Plumlee (2017)
Bayesian Calibration of Inexact Computer Models, *Journal of the American Statistical Association*