Calibration

STAT8810, Fall 2017

M.T. Pratola

November 29, 2017

Today

Combining Theoretical Models
and Observational Data
in a Probabilistic Framework
for Inference and Prediction.

Computer Model Calibration Experiments (CMCE's)

- A non-intrusive approach to combining simulator outputs, $\eta(\mathbf{x}, \mathbf{t})$ and observational ("field") data.
- Usually our simulator is expensive, so we have limited outputs we can run.
- And field data, $y^f(\mathbf{x})$ may be even more expensive, or otherwise difficult to obtain. Therefore, even fewer field observations.
- Here $\mathbf{x}$ are our usual control input variables as we saw when emulating. These are inputs that are also present for the observational data.

Computer Model Calibration Experiments (CMCE's)

- Simulators also typically depend on additional parameters, $\mathbf{t}$.
  - e.g. gravity in our ball-drop experiment
  - e.g. combustion parameter in our CO2 emissions problem.
- The simulator is linked to the real-world process through these unknown parameters, called $\{$*calibration parameters*$\}$.
- *Goal is to estimate* $\hat{\mathbf{t}} = \boldsymbol{\theta}$*, the parameter setting correponding to the real-world process.*
- *And predict the field process,* $y^f(\mathbf{x})$ *at new settings of* $\mathbf{x}$*, quantify uncertainties, etc.*
- *What if the simulator model is wrong? We can possibly estimate this discrepancy between the simulator and reality, called* $\boldsymbol{\delta}(\mathbf{x})$*, as well.*

CMCE Model†

- Our model for the field observations is

$$y^f(\mathbf{x}_i) = \eta(\mathbf{x}_i, \boldsymbol{\theta}) + \delta(\mathbf{x}_i) + \epsilon(\mathbf{x}_i), \quad i = 1, \ldots, n$$

where $\epsilon(\mathbf{x}_i) \sim N(0, \lambda_f^{-1})$, $\delta(\mathbf{x}_i)$ accounts for the discrepancy between the simulator and reality and $\boldsymbol{\theta}$ denotes the "true"(or best in some sense) setting of the calibration parameter $\mathbf{t}$.

† M.A. Kennedy and T. O'Hagan: {*Bayesian Calibration of Computer Models (with discussion)*}, *Journal of the Royal Statistical Society, Series B, vol.68, pp.425–464 (2001).*

CMCE Model, no discrepancy $(\delta(\mathbf{x}) = 0)$.

- Besides our model for the observations, we also need a model for the simulator outputs.
  - Since the simulator is slow, we will have to emulate it.
- We have field data,

$$\mathbf{y}^f = (y^f(\mathbf{x}_1), \ldots, y^f(\mathbf{x}_n))^T$$

- And simulator output,

$$\mathbf{y}^c = (y^c(\mathbf{x}_1, \mathbf{t}_1), \ldots, \mathbf{x}_m, \mathbf{t}_m)^T$$

- With no discrepancy, our model for the field is

$$y^f(\mathbf{x}_i) = \eta(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i$$

  and our model for the simulator is

$$y^c(\mathbf{x}_i, \mathbf{t}_i) = \eta(\mathbf{x}_i, \mathbf{t}_i)$$

CMCE Model, no discrepancy $(\delta(\mathbf{x}) = 0)$.

- Use our usual emulator model for the simulator, a GP:

$$\eta(\mathbf{x}, \mathbf{t}) \sim GP(\mu(\mathbf{x}, \mathbf{t}), \lambda^{-1}\mathbf{R}(\mathbf{x}, \mathbf{t}; \boldsymbol{\rho}))$$

where $\mathbf{R}(\mathbf{x}, \mathbf{t}; \boldsymbol{\rho})$ is formed as

$$\text{cor}(\eta(\mathbf{x}, \mathbf{t}), \eta(\mathbf{x}', \mathbf{t}')) = \prod_{i=1}^{d} c(\mathbf{x} - \mathbf{x}') \prod_{j=1}^{k} c(\mathbf{t} - \mathbf{t}')$$

for $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{t} \in \mathbb{R}^k$.

- A typical choice for the correlation function $c()$ will be the Gaussian.

CMCE Model, no discrepancy ($\delta(\mathbf{x}) = 0$).

- This gives us our model (and correspondingly the likelihood) for the field and simulator data,

$$\begin{pmatrix} \mathbf{y}^f \\ \mathbf{y}^c \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(\mathbf{x}, \boldsymbol{\theta}) \\ \mu(\mathbf{x}, \mathbf{t}) \end{pmatrix}, \lambda^{-1} \begin{bmatrix} \mathbf{R}^{ff} & \mathbf{R}^{fc} \\ \mathbf{R}^{cf} & \mathbf{R}^{cc} \end{bmatrix} + \begin{bmatrix} \lambda_f^{-1}\mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \right)$$

Here, $\mathbf{R}^{ff}$ denotes the correlation elements between field observations, $\mathbf{R}^{cc}$ the correlation between simulator outputs and $\mathbf{R}^{fc}$ the cross-correlation between field observations and simulator outputs.

CMCE Model, no discrepancy $(\delta(\mathbf{x}) = 0)$.

- For simplicity let's take $mu(\mathbf{x}, \mathbf{t}) = 0$.
- Specifying priors on the parameters $\boldsymbol{\rho}, \lambda, \lambda_f$ and the calibration parameters $\boldsymbol{\theta}$ we have

$$\pi(\boldsymbol{\theta}, \lambda, \lambda_f, \boldsymbol{\rho}|\mathbf{y}^f, \mathbf{y}^c) \propto L(\cdot|\mathbf{y^f}, \mathbf{y^c})\pi(\lambda)\pi(\lambda^f) \prod_{i=1}^{k} \pi(\theta_i) \prod_{j=1}^{d+k} \pi(\rho_j)$$

- Taking the same approach as our Bayesian GP regression model,

$$\pi(\lambda) = \text{Gamma}(a, b)$$
$$\pi(\lambda^f) = \text{Gamma}(a_f, b_f)$$
$$\pi(\rho_j) = \text{Beta}(\alpha_j, \beta_j)$$

- And we also need a prior on the calibration parameters,

$$\pi(\theta_i) = \text{Unif}(0, 1)$$

(assuming the inputs are scaled to the unit hypercube).

CMCE Model, no discrepancy $(\delta(\mathbf{x}) = 0)$.

- What does this model do? Consider predicting the field process at a new location ${x}^* **(for a given θ).**
- **Let $\mathbf{c}^T =$ $(\mathbf{cov}(y^f(\mathbf{x}^*), y^f(\mathbf{x}_1)), \ldots, \mathbf{cov}(y^f(\mathbf{x}^*), y^f(\mathbf{x}_n)), \mathbf{cov}(y^f(\mathbf{x}^*), y^c(\mathbf{x}_1)), \ldots$**
- **Or in short-hand, $\mathbf{c}^T = (\mathbf{c}^f, \mathbf{c}^c)^T$.**
- **Then the mean of the conditional predictive distribution is**

$$
\begin{aligned}
E[y^f(\mathbf{x}^*)|\mathbf{y}^f, \mathbf{y}^c, \cdot] &= \mathbf{c}^T \Sigma^{-1} (\mathbf{y}^f, \mathbf{y}^c)^T \\
&= \vdots \\
&= \sum_{i=1}^{n} w_i^f(\boldsymbol{\theta}) y^f(\mathbf{x}_i) + \sum_{j=1}^{m} w_j^c(\boldsymbol{\theta}) y^c(\mathbf{x}_j, \mathbf{t}_j)
\end{aligned}
$$

CMCE Model, no discrepancy $(\delta(\mathbf{x}) = 0)$.

- This shows that the field process is predicted as a weighted combination of the field observations and simulator output.
- The role of the estimated calibration parameter, $\boldsymbol{\theta}$, comes through the cross-covariance terms, $\mathbf{c}^c$ and Σ^{cf} which both depend on $\boldsymbol{\theta}$.
- If the estimated $\boldsymbol{\theta}$ indicates the field data is "far" from the simulator output, i.e. $|\theta_j - t_j|$ is large $\forall j$, then these correlation components will be small and the field prediction is mainly based on the field observations.
  - In extreme case of $\mathbf{c}^c = 0$ and $\Sigma^{cf} = 0$ we get $E[y^f(\mathbf{x}^*)] = {\mathbf{c}^f}^T {\Sigma^f}^{-1} \mathbf{y}^f$, the usual GP predictor.
- If the estimate of $\boldsymbol{\theta}$ is poor, the prediction of the field process may be inappropriately influenced by the simulator outputs if they receive too much weight – i.e. model things the outputs and field are "closer" than the actually are.

CMCE Model, with discrepancy

- Popular form of discrepancy is to assume an additive discrepancy,

$$y^f(\mathbf{x}_i) = \eta(\mathbf{x}_i, \boldsymbol{\theta}) + \delta(\mathbf{x}_i) + \epsilon_i$$

- Naturally, we will model the discrepancy, $\boldsymbol{\delta} = (\delta(\mathbf{x}_i), \ldots, \delta(\mathbf{x}_n))$ also as a GP,

$$\boldsymbol{\delta} \sim N\left(\mu_\delta(\mathbf{x}), \lambda_\delta^{-1}\mathbf{R}_\delta(\mathbf{x}; \phi)\right)$$

- Assuming η, δ and ϵ are independent, the likelihood becomes

$$\begin{pmatrix} \mathbf{y}^f \\ \mathbf{y}^c \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(\mathbf{x}, \boldsymbol{\theta}) + \mu_\delta(\mathbf{x}) \\ \mu(\mathbf{x}, \mathbf{t}) \end{pmatrix}, \Sigma\right)$$

where

$$\Sigma = \lambda^{-1} \begin{bmatrix} \mathbf{R}^{ff} & \mathbf{R}^{fc} \\ \mathbf{R}^{cf} & \mathbf{R}^{cc} \end{bmatrix} + \begin{bmatrix} \lambda_\delta^{-1}\mathbf{R}_\delta & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \lambda_f^{-1}\mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$$

MCMC Algorithm

- Similar to the one we developed for the Bayesian GP regression model.
- Except now we have a lot of parameters requiring Metropolis-Hastings steps.
- And we also need to sample the $\boldsymbol{\theta}$'s (MH as well).

Prediction and Inference

- We are typically interested in:
  - the emulated calibrated simulator, $E[\eta(\mathbf{x}, \boldsymbol{\theta}) | \mathbf{y}^f, \mathbf{y}^c]$
  - the predicted discrepancy, $E[\delta(\mathbf{x}) | \mathbf{y}^f, \mathbf{y}^c]$
  - the predicted field process, $E[\eta(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) | \mathbf{y}^f, \mathbf{y}^c]$
  - the estimated calibration parameter, $E[\boldsymbol{\theta} | \mathbf{y}^f, \mathbf{y}^c]$
- And of course uncertainties in the above.
- There are other forms of discrepancy that have been considered, such as multiplicative and more complex forms, but these are generally less common.