Calibration of computer models

Extension to Stochastic Simulator

Pierre BARBILLON

Fall 2023

Outline

1. Statistical Models
2. Heteroskedastic GP
3. Calibration
   - KOH
   - ABC

Stochasticity in Computer Experiments

The following is a basic model of a stochastic simulator experiment. If the code is run at a (vector) input x producing a (scalar) output $y(x)$, this could be represented as:

$$y(x) = M(x) + v,\ v \sim N(0, \sigma_v^2(x)), \tag{1}$$

where $M(x)$ is the expected value, $E[y(x)]$, of the output and v is independent variability representing the randomness of the simulator. Its variance, σ_v^2, can depend on x, but constant variance is also possible. For deterministic simulators, $\sigma_v^2 = 0$.
Stochasticity as a mean of computation
het or hom Stochastic

$$y_F(x) = y_S(x, u_C) + \delta_{\mathrm{MD}}(x) + \epsilon, \tag{2}$$

where $y_F(x)$ are real-world field observations at controllable (or measurable) inputs x, y_S is the simulator with additional unknown, non-measurable, inputs u_C, ϵ is measurement error for the observations $y_F(x)$ (with variance σ_ϵ^2), and $\delta_{\mathrm{MD}}(x)$ is an important term that accounts for the simulator not being a perfect representation of reality. y_F "observes" reality with error ϵ; reality = $y_S + \delta_{\mathrm{MD}}$.
For stochastic problems, where reality is stochastic, the discrepancy term $\delta_{\mathrm{MD}}(x)$ cannot be assumed deterministic. Discrepancy in stochastic settings is an open research question, with little attention so far. The model for the discrepancy may need to be similar to the model for the simulator; for example, if modeling y_S calls for a hetGP with a Matérn 5/2 correlation function then it is possible that a hetGP is needed for the discrepancy as well (perhaps with a smoother squared exponential correlation). A full Bayesian analysis in such circumstances may be prohibitively expensive and the above procedure might need to be modified. [Sung et al., 2019] use a hetGP for the discrepancy (but with a deterministic simulator), estimating parameters via maximum likelihood and following [Tuo et al., 2015] to avoid confounding.

Outline

1 Statistical Models

2 Heteroskedastic GP

3 Calibration
- KOH
- ABC

Stochastic Kriging

Observation model:

$$y_i^c = f(\mathbf{x}_i) + \epsilon_i, \quad \text{with} \quad \epsilon_i \overset{ind}{\sim} \mathcal{N}(0, r(x_i))\,.$$

In homoskedastic cases $r(x_i) = \tau^2$ which is called the nugget.
Stochastic Kriging provides for a design with replications These make up the "full-N" dataset, n of unique x_i-values in X_N with $n << N$, a_i replicates at unique locations and

$$\bar{y}_i = \frac{1}{a_i}\sum_{j=1}^{a_i} y_i^{(j)} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{a_i - 1}\sum_{j=1}^{a_i}(y_i^{(j)} - \bar{y}_i)^2.$$

the following BLUP:

$$\begin{aligned} \mu_n^{\mathrm{SK}}(x) &= \nu k_n^\top(x)(\nu C_n + S_n)^{-1}\bar{Y}_n \\ \sigma_n^{\mathrm{SK}}(x)^2 &= \nu K_\theta(x,x) - \nu^2 k_n^\top(x)(\nu C_n + S_n)^{-1} k_n(x), \end{aligned} \tag{10.1}$$

$k_n(x) = (K_\theta(x, \bar{x}_1), \ldots, K_\theta(x, \bar{x}_n))^\top$ $S_n = [\hat{\sigma}_{1:n}^2]A_n^{-1} = \mathrm{Diag}(\hat{\sigma}_1^2/a_1, \ldots, \hat{\sigma}_n^2/a_n)$, and
$C_n = \{K_\theta(\bar{x}_i, \bar{x}_j)\}_{1\leq i,j\leq n}$
[Ankenman et al., 2010]

as in [Goldberg et al., 1997] Heteroskedastic GP modeling assumes $\log(r(x)) \sim GP$ with zero mean and variance While full details are provided by [Binois et al., 2018], some specifics of the description above are worth noting. With $\lambda(x) = \sigma_V^2(x)/\sigma_Z^2$ and Λ_n = $(\lambda(x_1), \ldots, \lambda(x_n))$ for the n distinct inputs, $\log \Lambda_n$ is taken to be the predictive mean of a GP on latent (hidden) variables, $\Delta_n = (\delta_1, \ldots, \delta_n)$. For ease of exposition assume the GP has 0-mean (a constant mean is actually the default setting in `hetGP`) and take the covariance function for Δ_n to be $\sigma_g^2(C_g + gR^{-1})$ where $g > 0$, $R = \mathrm{diag}(r_1, \ldots, r_n)$, and C_g is a correlation function with parameters θ_g. Then $\log \Lambda_n = C_g(C_g + gR^{-1})^{-1}\Delta_n$. This latent Δ_n approach facilitates smooth estimates of Λ_n and provides a fixed functional form for $\lambda(x)$, but does not incorporate the resulting uncertainty. Given Λ_n, the Woodbury identities reduce the likelihood of Y_N, the output at all inputs including replicates, to depend only on quantities of size n. Maximum likelihood estimates for the unknown parameters can then be computed at a cost of $O(n^3)$, as can derivatives further facilitating optimization for maximizing likelihood.

Outline

1. Statistical Models

2. Heteroskedastic GP

3. Calibration
   - KOH
   - ABC

Outline

1 Statistical Models

2 Heteroskedastic GP

3 Calibration
- KOH
- ABC

Calibration of Stochastic Simulators

$$y_{exp_i} = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \delta(\mathbf{x}_i) + \epsilon(\mathbf{x}_i)\,. \tag{3}$$

$\delta(\cdot)$ models the difference between the simulator and the physical system:

$$\delta(\mathbf{x}) = \zeta(\mathbf{x}) - f(\mathbf{x}, \theta^*)\,.$$

Here *f* is Stochastic but its link with reality is questionable. Is reality $\mathbb{E}(f)$ or f? Depending on that, δ should be considered as deterministic or Stochastic and then modeled as a standard GP...

ocean example see https://github.com/Demiperimetre/Ocean

ref on HM

[**?**] contains a thorough description of HM whilst applying it to a complex epidemiology model of HIV.

Outline

1 Statistical Models

2 Heteroskedastic GP

3 Calibration
- KOH
- ABC

basics

ABC is a general method for producing samples from $\pi(u_C|Y_F)$, the posterior distribution of unknowns u_C, given data Y_F. ABC does this by generating samples for the unknowns $u_C^{(s)}$ and the output $z^{(s)}$ from $\pi(Y_F|u_C)\pi(u_C)$, that is, from the likelihood of the data given the unknowns, multiplied by the prior probability of the unknowns. For computer models, generating samples from the likelihood is equivalent to running the simulator. Such samples are only accepted if $z^{(s)} = Y_F$. For continuous settings, where exact equality cannot occur, acceptance is instead made if $B(z^{(s)}, Y_F) < \tau$, where B is a measure of distance and τ a level of tolerance. An approximated posterior distribution is then given by the collection of accepted $u_C^{(s)}$s. When there are multiple outputs (or there are other controllable inputs x, and so for any given $u_C^{(s)}$ there are effectively multiple outputs), Y_F and $z^{(s)}$ can be replaced with informative summary statistics. Finding a single statistic sufficient for all outputs is challenging, and a poorly chosen one can invalidate results.

The choice of the tolerance τ is also important. If τ is small then it may take a very long time to generate a single sample which satisfies the inequality. If τ is not small then the approximation to the posterior is less reliable. For calibration, τ can be interpreted as a bound on the observational error and model discrepancy, leading to a “correct” posterior rather than an approximation [Wilkinson, 2013]. This is then similar to HM with the subjective choice of bounds.

ABC can be done without the use of a surrogate, but this will require many runs of the

fish example see
https://github.com/jhuang672/fish/blob/master/fish_fits.md

Ankenman, B., Nelson, B. L., and Staum, J. (2010).
Stochastic kriging for simulation metamodeling.
Operations Research, 58(2):371–382.

Binois, M., Gramacy, R. B., and Ludkovski, M. (2018).
Practical heteroscedastic Gaussian process modeling for large simulation experiments.
Journal of Computational and Graphical Statistics, 27(4):808–821.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1997).
Regression with input-dependent noise: a Gaussian process treatment.
In Proceedings of the 10th International Conference on Neural Information Processing Systems, pages 493–499.

Sung, C.-L., Barber, B. D., and Walker, B. J. (2019).
Calibration of computer models with heteroscedastic errors and application to plant relative growth rates.
arXiv preprint arXiv:1910.11518.

Tuo, R., Wu, C. J., et al. (2015).
Efficient calibration for imperfect computer models.
The Annals of Statistics, 43(6):2331–2352.

Wilkinson, R. D. (2013).

Approximate Bayesian computation (abc) gives exact results under the assumption of model error.
Statistical Applications in Genetics and Molecular Biology, 12(2):129–141.