Calibration of computer models

Extension to Stochastic Simulator

Pierre Barbillon

Fall 2023

Outline

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

Outline

Stochasticity in Computer Experiments

Basic model for a stochastic simulator:

$$f(\mathbf{x}) = m(\mathbf{x}) + v, \; v \sim N(0, \sigma_v^2(x)), \tag{1}$$

where

- $m(\mathbf{x})$ is the expected value: $\mathbb{E}_f[f(\mathbf{x})]$,
- v independent variability representing randomness of the simulator,
- variance σ_v^2 may depend on $\mathbf{x}$, be constant.

Remarks

- If $\sigma_v^2 = 0$ deterministic simulator,
- What does stochasticity acknowledge for? numerical approximation (Monte Carlo), aleatory experiment.

Extension of KOH

$$y_i^e = y^e(\mathbf{x}_i^e) = f(\mathbf{x}_i^e, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^e) + \epsilon_i, \quad (2)$$

where

- $\mathbf{y}^e = \{y_1^e, \ldots, y_{n_e}^e\}$ are real-world field observations at controllable (or measurable) inputs $(\mathbf{x}_i)_{1 \leq \ldots \leq n_e}$,
- f is a stochastic simulator,
- ϵ is measurement error for the observations,
- δ is the discrepancy may be assumed to be stochastic.

Remarks: If reality is stochastic, δ has to be stochastic and may be heteroskedastic as the simulator.
y_S is the simulator with additional unknown, non-measurable, inputs u_C, $y_F(x)$ (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 = $y_S + \delta_{\mathrm{MD}}$.

[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

Stochastic Kriging

[Ankenman et al., 2010]
Observation model:

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

In homoskedastic cases $r(\mathbf{x}_i) = \tau^2$ which is called the nugget.

From a design with replications:

- 'full-N" dataset, n of unique x_i-values in X_N with $n << N$, a_i replicates at unique locations,
- compute

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

predictions with BLUP for M when GP is assumed on M:

$$\mu_n^{\mathrm{SK}}(\mathbf{x}) = k_n^\top(\mathbf{x})(C_n + S_n)^{-1}\bar{Y}_n$$
$$\sigma_n^{\mathrm{SK}}(\mathbf{x})^2 = c_{S,\psi}(\mathbf{x}, \mathbf{x}) -^2 k_n^\top(\mathbf{x})(C_n + S_n)^{-1}k_n(\mathbf{x}),$$

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

Modeling the variance

- In [Ankenman et al., 2010], no specific model for the variance,
- [Goldberg et al., 1997] assumes $\log(r(\mathbf{x})) \sim GP$ for modeling heteroskedasticity, and they estimate the combined parameters of the two GPs with an MCMC scheme.

[Binois et al., 2018] make use of Stochastic Kriging with GP model for $\log(r(\mathbf{x}))$.

- consider latent variances: $\xi_1, \ldots, \xi_n$ for the n unique locations,
- GP prior on this matrix $\Xi_n \sim \mathcal{N}_n(0, \nu(C_\xi + g_\xi A_n^{-1}))$ where g_ξ regularizes the behavior of the variance process,
- Estimate parameters by MLE using Woodbury trick which put all the computation in $\mathcal{O}(n^3)$,
- implementation in `hetGP` package.

homoskedastic GP

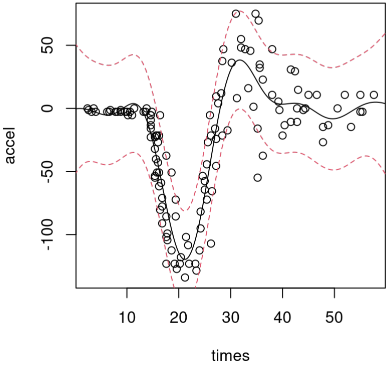

Figure: Homoskedastic GP fit to the motorcycle data via mean (solid-black) and 90% error-bars (dashed-red).

from [Gramacy, 2020]

heteroskedastic GP

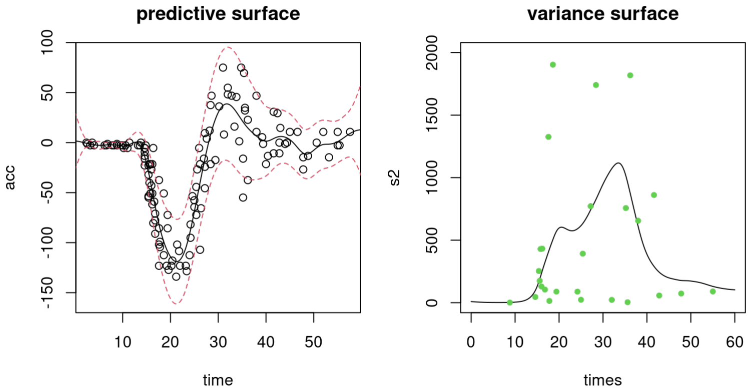

Figure: Heteroskedastic GP fit to the motorcycle data. Left panel shows the predictive distribution via mean (solid-black) and 90% error-bars (dashed-red). Right panel shows the estimated variance surface and moment-based estimates of variance (green dots).

from [Gramacy, 2020]

Outline

Outline

Calibration of Stochastic Simulators

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

$\delta(\cdot)$ 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

History Matching

[Andrianakis et al., 2015] contains a thorough description of HM whilst applying it to a complex epidemiology model of HIV.

Outline

Basics

Approximate Bayesian Computation produces samples from a posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}^e)$ by

- generating samples for $\boldsymbol{\theta}$ from the prior
- and outputs $\mathbf{y}$ from the generating model $\pi(\cdot|\mathbf{y})$ (this implies runs of the simulator),
- samples are kept provided that $\mathbf{y} = \mathbf{y}^e$ or $|h(\mathbf{y}) - h(\mathbf{y}^e)| < tol$,
- accepted $\boldsymbol{\theta}$s produce an approximated posterior sample.

Remark:

- For calibration, *tol* 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 simulator itself. Otherwise, very few accepted $\boldsymbol{\theta}$ will be obtained, or an overly high value of *tol* will be required.

Fish example

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

Other papers

[Oakley and Youngman, 2017] removes δ but compensates by inflating the variability in the simulator output.

Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2015).
Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on hiv in uganda.
PLoS computational biology, 11(1):e1003968.

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.

Gramacy, R. B. (2020).
Surrogates: Gaussian process modeling, design, and optimization for the applied sciences.
CRC press.

Oakley, J. E. and Youngman, B. D. (2017).
Calibration of stochastic computer simulators using likelihood emulation.
Technometrics, 59(1):80–92.

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.