Bayesian Inference Using Gaussian Process Metamodel

Colin Cui

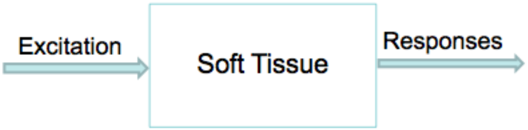
May 8, 2016

Outline

Background information

Biomechanical imaging

- Imaging soft tissues based on their mechanical properties
- The excitation: surface or internal, static or dynamic
- The responses: strain, displacement, phase angle
- Imaging: strain imaging, stiffness contrast or absolute mechanical properties



Input-output system representation of soft tissue investigation

- black-box, inverse problem

Bayesian inference

Bayesian inference

1. Bayes' rule

$$p(\theta|D) = \frac{p(D|\theta) \times p(\theta)}{p(D)}$$

$$p(D) = \int p(D|\theta)p(\theta)d\theta$$

The posterior is normalized by $p(D)$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(\theta|D) \propto p(D|\theta) \times p(\theta)$$

2. Posterior: intractable integral
   - MCMC, variational inference
   - Maximum-a-Posteriori

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$

Gaussian Process I

Gaussian Process $y = f(t) + \epsilon$

$$f(t) \sim N(m(t), K(t, t'))$$

- mean function: $m(t)$
- kernel function: $K(t, t') = \sigma_f^2 exp\{-\frac{1}{2}\sum_{i=1}^{d}(t - t')^2/l_i\}$
  - prior: $\sigma_f^2, l \sim p(\sigma_f^2)p(l)$
  - hyperparameters: σ_f^2, l

A random GP random vector,

$$y \sim N(\mathbf{0}, K(X, X))$$

Thus,

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim N\left(\mathbf{0} , \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

Gaussian Process II

Gaussian Process is a latent function, we will integrate it out

$$p(y|X) = \int p(y|f,x)p(f|X)df$$

$$\log p(y|X) = -\frac{1}{2}y^T(K+\sigma_f^2 I)^{-1} - \frac{1}{2}\log|K+\sigma_f^2 I| - \frac{n}{2}\log 2\pi$$

This result can be observed directly by $y \sim N(0, K+\sigma_f^2 I)$

Bayeisan Calibration: Overview

The Bayesian calibration model used in this study,

Full model

$$y^e = y^m(x, \boldsymbol{\theta}) + \delta(x) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, {\sigma_\epsilon}^2)$

- $y^m(x, \boldsymbol{\theta})$: computer model
  - x: input design
    - for example, latin hypercube
  - $\boldsymbol{\theta}$: parameter
- $\delta(x)$: discrepancy function
  - bias of the computer model
- y^e: observation
- black-box, true model unknown

Bayesian Calibration: the computer model

Computer model: y=$\eta(X, \boldsymbol{\theta}) + \epsilon$

- Computer model $y(x, \boldsymbol{\theta}, ee)$ input
  - x: input design location
  - $\boldsymbol{\theta}$: **estimation parameter** time constant
  - ee: elasticity
- y: model output

Full model: $z = \eta(x, \boldsymbol{\theta}) + \delta(x) + \epsilon$

1. Observation points
   - x: observation input
   - z: observation output
2. Throwing basketball into a well, let it bounce back

GP for Computer Model

Full model

$$z = \eta(x, \boldsymbol{\theta}) + \delta(x) + \epsilon$$

Part 1: Computer model using GP

Computer model

$$y = \eta(x, \boldsymbol{\theta}) + \epsilon$$

$$\eta(x, \theta) \sim N(0, K\{(x, \theta), (x, \theta)\})$$

$K(x, \theta), (x, \theta) =$

$$\sigma_\epsilon^2 \exp\{-(x - x')^T \Omega_x (x - x')\} \exp\{-(\theta - \theta')^T \Omega_\theta (\theta - \theta')\}$$

where Ω_x and Ω_θ are diagonals
Hyperprior: $\sigma_\epsilon^2, l_1 \sim p(\sigma_\epsilon^2) p(l_1)$
Optimize: σ_ϵ^2, l_1

Bayesian Calibration: Discrepancy Function

Full model

$$z = \eta(x, \boldsymbol{\theta}) + \delta(x) + \epsilon$$

Part 2: discrepancy function using GP

Discrepancy function

$$\delta(x) = z(x) - y(x, \boldsymbol{\theta})$$

$$\delta(x) \sim N(0, K(x, x))$$

GP for discrepancy function

$\delta(x)$ is modeled as a GP

- discrepancy function aims to improve any lacking of the computer model
- unknown but true θ at $y^m(x, \theta)$

$$\delta(x) = y^e(x) - y^m(x, \theta)$$

- build GP for discrepancy function $\delta(x)$

$$\delta(x) \sim N(0, K(x, x'))$$

$$K(x, x') = \sigma_\delta^2 \exp\{-(x - x')^T \Omega_x (x - x')\}$$

Hyperprior: $\sigma_\delta^2, l_2 \sim p(\sigma_\delta^2)p(l_2)$
Optimize: σ_δ^2, l_2

The Bayesian framework

Full Bayesian treatment

Full Bayesian posterior

$$p(\theta, \phi | D) \propto p(\theta)p(\phi)f(D; m, V)$$

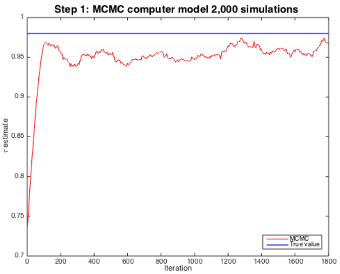
where $\phi = \{\sigma_\epsilon^2, \sigma_\delta^2, l_1, l_2\}; f \sim N(m, V)$

- information theory perspective
  - the amount of information gain after observing the data
- considers all sources of uncertainty
  1. parameter uncertainty: θ
  2. latent model uncertainty: y^m
  3. obervation uncertainty: y^e
  4. hyperparameter uncertainty: ϕ

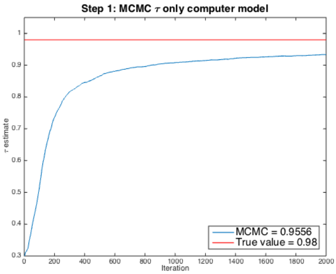
Results: computer model only

Computer model only, and estimate τ

- prior: $\tau \sim N(0.8, 0.6^2)$
- $y^e = y^m(\mathbf{x}, \tau) + \epsilon$
- posterior: $\tau \sim N(0.9556, 0.0689^2)$
- MCMC 2000 simulations
- computer model alone is not sufficient to converge



Computer model random walk

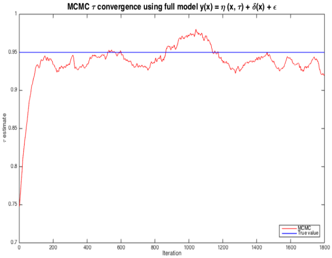


Computer model convergence

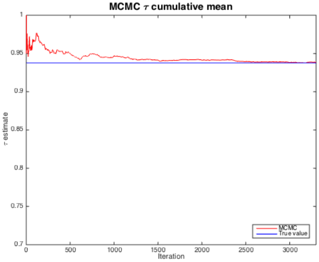
Results: Full model

Estimate τ using full model

- prior: $\tau \sim N(0.8, 1)$
- model: $y^e = y^m(\mathbf{x}, \tau) + \delta(\mathbf{x}) + \epsilon$
- posterior: $\tau \sim N(0.9813, 0.0969^2)$
- MCMC 10000 simulations



MCMC moving average (simulation is stable)



MCMC convergence (algorithm converged)

Parametric study: compare priors I

Uninformative prior

2000 simulations

τ	Uniform
prior	U(0.3, 1.3)
posterior	$N(0.9766, 0.0983^2)$

10000 simulations

τ	Uniform
prior	U(0.3, 1.3)
posterior	$N(0.9825, 0.1003^2)$

Informative prior

2000 simulations

τ	mean	sd
prior	1	0.6
posterior	0.9795	0.0992

Large sd, same run, conv.

2000 simulations

τ	mean	sd
prior	1	0.2
posterior	0.9855	0.0897

Small sd, same run, almost conv.

Parametric study: compare priors II

Informative prior

5000 simulations

τ	mean	sd
prior	1.8	0.6
posterior	1.0085	0.1037

Bigger sd, short run, no conv.

20000 simulations

τ	mean	sd
prior	1.8	0.2
posterior	1.16	0.1148

Smaller sd, long run, no conv.

Informative prior

2000 simulations

τ	mean	sd
prior	1	0.6
posterior	0.9795	0.0992

Large sd, same run, conv.

2000 simulations

τ	mean	sd
prior	1	0.2
posterior	0.9855	0.0897

Small sd, same run, almost conv.

Which one is better: uniform vs biased mean with small sd

Choices of prior: uniform vs. normal biased mean and small sd

- uninformative prior: $\tau \sim U(0.3, 1.3)$
- informative prior, but biased prior: $\tau \sim N(1.8, 0.2^2)$
- Let both chains run for long time
- true τ estimate 0.98

Compare priors

10000 simulations

τ	Uniform
prior	U(0.3, 1.3)
posterior	$N(0.9825, 0.1003^2)$

This performed better

10000 simulations

τ	Normal
prior	$N(0.8, 0.2^2)$
posterior	$N(0.9460, 0.0915^2)$

Conclusion

Why Bayesian inferece?

- parameter of interest not fixed
- impose prior belief on the unknown
- includes all sources of uncertainty

Why GP?

- advantages of GP over other choices
  - metamodel: model about a model
  - a flexible class of function, takes shape of data
  - a nonparametric function
  - a latent fuction, to be integrated out, not used for estimation

Reference

- Kennedy, Marc C and O'Hagan, Anthony. *Bayesian calibration of computer models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), pp. 425-464.
- Zhao, X and Pelegri, AA. *A Bayesian approach for characterization of soft tissue viscoelasticity in acoustic radiation force imaging*. Int. J. Numer. Meth. Biomed. Engng, (2015).
- Rasmussen, C. E., and Williams, C. K. I., 2005, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press.

Thank you

Thank you! Any Questions?